Molecular Latent Space Simulators for Distributed and Multi-Molecular Trajectories

Michael S. Jones,[†] Zachary A. McDargh,[‡] Rafal P. Wiewiora,[‡] Jesus A. Izaguirre,[‡] Huafeng Xu,[‡] and Andrew L. Ferguson[*,†]

† *Pritzker School of Molecular Engineering, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA*

‡ *Roivant Discovery, 151 W 42nd Street, New York, New York 10036, USA*

E-mail: andrewferguson@uchicago.edu

Abstract

All atom molecular dynamics (MD) simulations offer a powerful tool for molecular modeling, but the short time steps required for numerical stability of the integrator place many interesting molecular events out of reach of unbiased simulations. The popular and powerful Markov state modeling (MSM) approach can extend these timescales by stitching together multiple short discontinuous trajectories into a single long-time kinetic model, but necessitates a configurational coarse-graining of the phase space that entails a loss of spatial and temporal resolution and an exponential increase in complexity for multi-molecular systems. Latent space simulators (LSS) present an alternative formalism that employs a dynamical, as opposed to configurational, coarse graining comprising three back-to-back learning problems to (i) identify the molecular system's slowest dynamical processes, (ii) propagate the microscopic system dynamics within this slow subspace, and (iii) generatively reconstruct the trajectory of the system within the molecular phase space. A trained LSS model can generate temporally and spatially continuous synthetic molecular trajectories at orders of magnitude lower cost than MD to improve sampling of rare transition events and metastable states to reduce statistical uncertainties in thermodynamic and kinetic observables. In this work, we extend the LSS formalism to short discontinuous training trajectories generated by distributed computing and multi-molecular systems without incurring exponential scaling in computational cost. First, we develop a distributed LSS model over thousands of short simulations of a 264-residue proteolysis-targeting chimera (PROTAC) complex to generate ultra-long continuous trajectories that identify metastable states and collective variables to inform PROTAC therapeutic design and optimization. Second, we develop a multi-molecular LSS architecture to generate physically realistic ultra-long trajectories of DNA oligomers that can undergo both duplex hybridization and hairpin folding. These trajectories retain thermodynamic and kinetic characteristics of the training data, while providing increased precision of folding populations and timescales across simulation temperature and ion concentration.

1 Introduction

The expanding capabilities of molecular dynamics (MD) simulations have motivated the development of many data-driven approaches to distill, interpret, and model increasing complex systems. Markov state models (MSMs) are a powerful and popular approach to constructing interpretable long-time kinetic models by reducing high-dimensional MD simulations into a set of discrete configurational states and modeling the system's kinetic evolution via transition probabilities between these states.[1,2] In the past decade, numerous techniques – increasingly integrating deep learning – have been introduced to optimize some aspect of the MSM pipeline[3–5] or to replace the conventional pipeline all together.[6,7] Although tremendous progress has been made in this area, MSMs inherently model jump processes between coarse-grained metastable states.[8] In addition to these spatial and temporal limits, global MSMs can be inefficient or misleading when applied to large or multi-component systems.[9]

Recently, latent space simulators (LSS) were developed as learned kinetic models capable of generating temporally continuous synthetic molecular trajectories at a fraction of the computational costs of MD.[10] The LSS pipeline is trained on MD training data and uses three distinct deep learning architectures to (i) encode molecular trajectories into a latent space (ii) propagate low-dimensional trajectories in that latent space and (iii) decode latent space back to configurational space. The fully trained pipeline learns the statistics of the infinitesimal generator of microscopic transition elements and can therefore produce physically realistic molecular trajectories that are distinct from training data while still reproducing thermodynamic and kinetic observables. The LSS approach shares similarities with the variable-free/equation-free approach of Kevrekidis and co-workers,[11–17] but employs the transfer operator formalism to directly estimate the slow modes and comes equipped with decoders to efficiently lift back up to the high-dimensional molecular phase space. Previous techniques for synthetic trajectory generation based on time-lagged autoencoders were limited by unstable propogators,[18] and MSM-based approach such a deep generative MSMs relied on a discretieztion of the latent space.[19] Recently several approaches[20–23] have been

introduced to generate synthetic trajectories using similar procedures to LSS, including learning of effective dynamics (LED)[23] which uses a recurrent structure to add long-term memory to the propagator. The LSS approach remains unique, however, in its ability to encode configurations into an optimally slow basis, learn a microscopic generator in the latent space, and decode the latent space using a high-fidelity generative model. Previously, the LSS was applied to the fast-folding Trp-cage mini-protein, where thermodynamically and kinetically accurate ultra-long simulations were generated at several orders of magnitude lower cost compared to MD.[10] In this work, we have extended the LSS pipeline to two applications which represent distinct challenges for kinetic modeling and synthetic trajectory generation: training over short, discontinuous trajectories and multi-molecular systems.

First, we explore the training of LSS models not over a single long continuous trajectory, but over a number of short MD simulations generated by distributed computing. In MSMs, the configurational coarse-graining into metastable states presents a relatively forgiving training paradigm since the learning problem is well posed provided there are sufficiently many pathways linking these states. The LSS eschews this configurational coarse-graining for a dynamical one into the leading slow modes. Short trajectories that are not guaranteed to chart a continuous path through phase space could conceivably frustrate learning of a slow subspace and/or the transition density elements within this projection. Training of LSS models in this mode is desirable for their application to large molecular systems where generation of multiple short trajectories using distributed computing is more computationally accessible than performing a single long calculation.

We demonstrate stable training of LSS models over distributed computing training data in an application to a ternary degrader complex recently interrogated by Dixon et al.[24] This system comprises two large (>100 residue) proteins joined by a proteolysis-targeting chimera (PROTAC) molecule. PROTACs compounds represent a novel and desirable drug design paradigm in which a ubiquitin ligase is recruited to the protein of interest to mark the latter for degradation.[25] As opposed to stoichiometric small molecule or biologic inhibitors, a

PROTAC compound can be recycled after each degradation event and so effectively target proteins at much lower concentration.[26] In the two decades since PROTACs were first demonstrated to facilitate degradation,[27] there has been tremendous progress in improving target specificity and translating PROTACs molecules into clinically promising therapeutics.[28] Computational investigations of ternary PROTAC complexes have, however, been relatively limited due in part to the system size and diversity of states involved in the binding mechanisms.[29–31] While enhanced sampling techniques can traverse large free energy barriers and improve thermodynamic understanding,[24] unbiased data is required to rigorously probe kinetics and binding mechanisms. In this work, we train the LSS on 9,800 independent and relatively short ∼650 ns equilibrium trajectories of the PROTACs system collected from distributed Folding@Home simulations.[32] This represents the first training of the LSS on distributed data, and we show that the LSS can effectively knit together thousands of independent trajectories comprising 5.7 ms of total simulation time to produce a unified long-time kinetic model. Similar to MSMs, the trained LSS model can produce synthetic trajectories with comparable structural and thermodynamic observables to the discontinuous training data at low computational cost, but unlike an MSM, the LSS trajectories are temporally and spatially continuous. The trained LSS model can produce a 260 ms synthetic trajectory in several GPU-minutes compared to the estimated ∼7000 GPU-years required to produce the same amount of data using Folding@Home. Furthermore, the LSS furnishes inexpensive continuous trajectories that can efficiently and densely sample rare events in the configurational phase space and, by learning the slowest dynamical modes across all training trajectories, furnishes interpretable slow physical processes and metastable states we have correlated with PROTAC degradation efficiency. These learned slow modes can thereby provide a basis to evaluate and optimize future PROTAC drug candidates by computational screening.

Second, we make several architectural additions to the LSS to render it applicable to multi-molecular systems. It has been shown MSMs can be inefficient, ineffective, or mis-

leading when they are applied to multi-molecular or partially coupled systems.[9,33] Recent work to address the shortcomings has been driven in large part by Noé and co-workers, and progress has been made in the development of so-called Markov field models.[9] This includes efforts to learn independent components of bio-molecular systems[33,34] and couple MSMs via reaction-diffusion dynamics.[35,36] Here, we integrate a novel approach, related to that of del Razo et al.,[36] into the LSS pipeline which enables us to independently encode, propagate, and decode molecular subsystems and generate ultra-long synthetic trajectories for multi-molecular systems. To avoid learning degenerate dynamics, we build separate encoders and latent spaces for each subsystem and use a joint propagator to ensure physically accurate coupling between each system. We train additional propagators on the translational and rotational degrees of freedom between strands to preserve the orientation of transition states and ensure no overlap between dissociated configurations. Finally, we convert latent space trajectories back to molecular configurational space either by decoding the complete system directly or by decoding each strand independently and reconstructing the system based on a learned inter-molecular orientation.

We demonstrate this multi-molecular generalization of the LSS approach (Multi-LSS) on coarse-grained trajectories of two DNA strands that can form both duplex and hairpins. The bimolecular LSS model generates stable and physically realistic synthetic molecular trajectories at five orders of magnitude lower cost than MD that preserve both the global dynamics of duplex folding and the single-strand dynamics of hairpin folding. We learn interpretable slow collective variables describing the dynamics of independent subsystems and leverage these modes to generate novel and physically meaningful synthetic trajectories which reproduce thermodynamic and kinetic observables with significantly lower uncertainties. Furthermore, we show that our encoder/decoder can be extended to MD simulations under new temperatures and ion concentrations without re-training the complete pipeline. The trained LSS models provide better understand the role of secondary structure on hybridization[37–41] and inform potential applications of these sequences in DNA nanotechnology.[42,43]

2 Methods

2.1 All-atom simulations of PROTAC ternary complex

The PROTAC complex comprises three components and is schematically illustrated in Figure 1: (i) the 115-residue bromodomain of the human SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2 (SMARCA2 or SMC2) protein, (ii) the the 149-residue von Hippel-Lindeu disease tumor suppressor protein (VHL), and (iii) a heterobifunctional small molecule proteolysis targeting chimera (PROTAC) known as PROTAC2 that bridges SMC2 and VHL to form the ternary complex (PDB ID: 6HAX). All-atom simulations of the complex were conducted as detailed in Ref.[24]. In brief, the ternary complex was solvated in explicit water and net charge was neutralized by the addition of counter ions. The Amber ff14SB force field[44] was chosen for proteins, the TIP3 water model[45] was used for solvent, and in-house force field parameters were generated for the PROTAC2 molecule. The Verlet leapfrog algorithm[46] was used integrate equations of motion at a time step of 2 fs, and the LINCS algorithm was employed to restrain hydrogen bonds.[47] Particle-mesh Ewald summation[48] was used to treat long-range electrostatic interactions employing a real-space cutoff of 1.2 nm. A velocity rescaling thermostat[49] and Parrinello-Rahman barostat[50] were used to maintain system at 310 K and 1 atm. Energy minimization was performed with the steepest descent algorithm, and equilibration was performed in the NVT and NPT ensembles. All production runs were performed in the NPT ensemble. Initial configurations for each short discontinuous simulation were seeded from Hamiltonian replica-exchanged MD, in which trajectories were exchanged between replicas containing altered potential energy functions in order to enhance sampling.[51] To ensure adequate coverage of the sampled phase space, k-means clustering was performed on a PCA projection of previously collected Hamiltonian replica exchange molecular dynamics simulation data. A total of 98 seeds were collected from k-means centers, and 100 independent simulations were performed from each seed. Simulations were distributed across Folding@Home[32] computing resources

to produce at total of 9800 trajectories with a median length of ∼650 ns for an aggregated total of ∼5.7 ms of simulation time. The estimated computational cost to produce these trajectories is ∼7000 GPU-years.

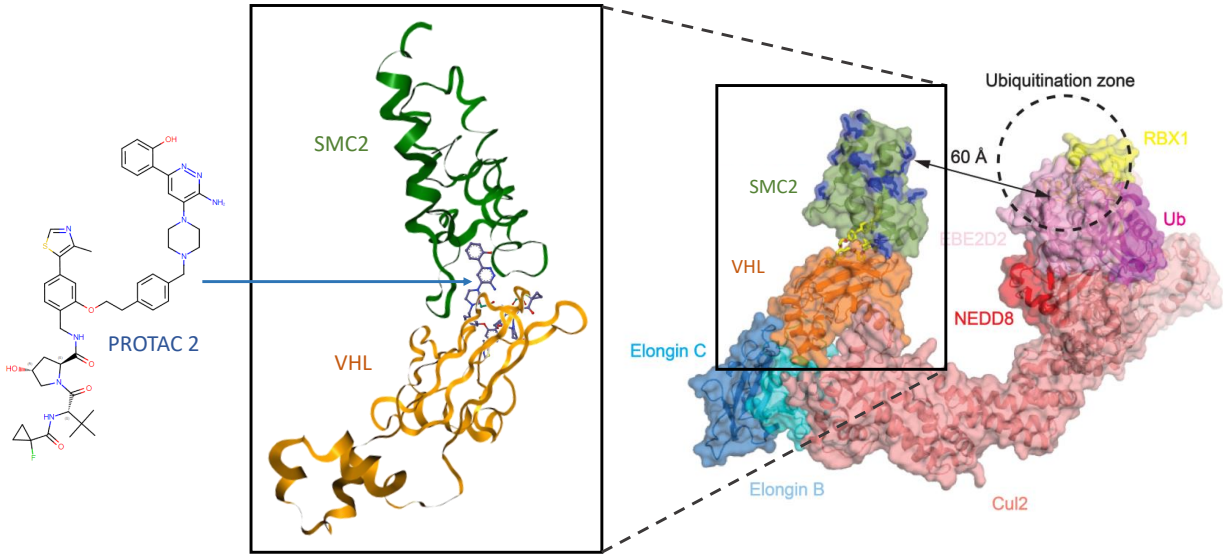

Figure 1: Schematic of the VHL-PROTAC2-SMC2 ternary complex structure (inset) and its placement within the full Cullin-RING E3 ubiquitin ligase (CRL). The inset presents ribbon diagrams of the SMC2 (green) and VHL (orange) proteins as well as the implicitly modeled PROTAC2 molecule (blue). The chemical structure of PROTAC2 is presented to the left. The space filling rendering of the full CRL complex illustrates how the VHL component of the ternary complex attaches to the Elongin C, Elongin B, Cul2, NEDD8, EBE2D2, Ub, and RBX1 proteins, and how the protein of interest, here SMC2, is placed within ∼60 Å of the ubiquitination zone enabling marking the target protein for proteosomal degradation. The image of the protein complex on the right is adapted from Dixon et al.[24] with permission under the Creative Commons Attribution 4.0 International License (`www.creativecommons.org/licenses/by/4.0/`) Copyright 2022 Springer Nature.

The aggregated simulation data were used to train an LSS model for the dynamical evolution of the PROTAC ternary complex. The trajectories were subsampled at a 5 ns stride and featurized by calculating the sin/cos of all backbone dihedral angles in addition to the inverse pairwise distances between every fourth protein C_α. These distances represent a set of intra-molecular distances within VHL and SMC2 as well as inter-molecular distances between

the two molecules. This procedure produced a total of 3197 translationally and rotationally invariant features that were normalized to a $[0, 1]$ range prior to passing to the LSS encoder. The LSS decoder was trained to reconstruct the aligned Cartesian coordinates of the alpha carbons of the SMC2 and VHL proteins from the latent space coordinates. The PROTAC2 small molecule was modelled implicitly by the LSS, therefore features of its geometry were not encoded into the latent space and it was not reconstructed during the decoding process. A consistent alignment was generated by performing Procrustes alignment to a metastable "hub" configuration (see Section 3.1) using the Kabsch algorithm.[52,53] implemented in the mdtraj package.[54]

2.2 Coarse-grained simulations of DNA duplex and hairpin formation

We conducted coarse-grained MD simulations of a duplex DNA system containing two identical strands (5′-GCGGTTTCCGC-3′) designed to admit both inter-molecular hybridization and intra-molecular folding into a hairpin. The system is capable of forming up to eight inter-molecular Watson-Crick-Franklin (WCF) pairs and four intra-molecular WCF pairs, where the duplex structure is destabilized by four internal T-T mismatches. Hairpin structures with GC-rich stems and poly-T loops are frequently used in DNA nanotechnology owing to the relative strength of G:C hydrogen bonds and the flexibility of consecutive thymine stretches.[55–57] We constructed and simulated the DNA sequences using the coarse-grained 3-Site-Per-Nucleotide v2 (3SPN.2) model that uses three spherical beads to represent the phosphate, deoxyribose sugar, and nitrogenous base of each nucleotide and employs anisotropic interaction potentials to accurately treat intra-strand base-stacking, inter-strand cross-stacking, and base pairing.[58] The model was parameterized against experimental structural and kinetics data and has been extensively validated against experimental comparison of hybridization mechanisms[59,60] and protein-DNA interactions.[61,62] Kinetic timescales have been shown to be accelerated compared to experiment, however, relative timescales can be

reliably evaluated after applying a corrective factor.[60]

All 3SPN.2 simulations were performed via the LAMMPS plugin accordance with best practices[58,63] Two identical strands were placed in a cubic periodic box with side length 8.5 nm corresponding to a single-strand concentration of 5.4 mM. Solvent effects were modeled implicitly by employing Langevin dynamics[64,65] with an experimentally motivated per-site friction coefficient of 9.94×10^{-11} m^2/s.[58,66] For wild type (WT) runs we specified a 100 mM implicit NaCl concentration and treated electrostatic interactions using the Debye-Hückel with a 5 nm cutoff radius.[67] Simulations were performed in the NVT ensemble employing a Langevin thermostat.[68] A simulation temperature of 320 K was selected to maintain a large and approximately equal population of both duplex and hairpins state and to maximize the number of hybridization and folding events. For the transfer learning procedure, ion concentrations were varied from 25 mM to 400 mM while holding temperature at 320 K; temperatures were then varied from 310 K to 330 K while holding implicit ion concentration at 100 mM. The Langevin equations of motion were integrated using the scheme of Bussi and Parrinello[65] with a 20 fs integration time step. We performed 10 independent simulations for each temperature and ion concentration with half of the runs initialized from the hybridized state and half from the dissociated state. The initial hybridized state was defined based on the crystal structure coordinates of Arnott et al.[69] The dissociated state was generated from the hybridized state by displacing one strand away from the other by 1 nm in each of the x, y, and z directions. Initial bead velocities were assigned from a Maxwell-Boltzmann distribution at the temperature of interest. Each simulation was conducted for 20 μs and frames saved to disc every 100 ps. Each simulation required $\sim$24 CPU-hours on 28$\times$Intel E5-2680v4 CPU cores. For WT simulations, we generated ten independent trajectories corresponding to a combined total of 200 μs simulation time, in which we sampled 18 hybridization events, 22 dehybridization events, and hundreds of hairpin folding and unfolding events. Based on a center-of-mass cutoff of 1.5 nm between strands, we found that 27% of configurations were in a hybridized state, 37% contained at least one folded hairpin, and 6% had both strands

folded into hairpins.

To train the LSS encoder, simulation data were subsampled at a 100 ps stride and transformed into a set of translationally and rotationally invariant features. For the global system comprising both DNA strands (double stranded, DS) we used all inter-molecular distances between nucleobases on each strand, corresponding to 121 total features. For the two single-stranded DNA subsystems (strand 1, S1, and strand 2, S2) we extracted all possible intra-molecular distances from each nucleobase for a total of 55 features each on each strand. Each set of distances was normalized to a $[0, 1]$ range prior to passing to the network. To train the LSS decoder, we employed the orthogonal Proscrutes solver[70] implemented in scipy[71] to align the DS and S1/S2 configurations to reference crystal structures.[69] For de-hybridized frames, the relative distance and angles transformations applied to superpose each strand were recorded for each frame and saved as three-dimensional translations and three-dimensional Euler angles. Together, these six-dimensional coordinates describe all translational and rotational degrees of freedom between S1 and S2.

2.3 Latent space simulators (LSS)

Molecular latent space simulators comprises three deep learning architectures to (i) encode molecular trajectories into a latent space, (ii) propagate low-dimensional trajectories in that latent space, and (iii) decode latent space back to configurational space. The low-dimensional embedding learned by the encoder is required for training the propagator and decoder. Therefore, once the encoder has been trained, the propagator and decoder can be trained independently and in parallel. The three back-to-back learning problems are independent and may be trained in a sequential fashion. The fully trained pipeline generates physically realistic molecular trajectories that are distinct from training data but still reproduce accurate thermodynamic and kinetics. Full details of the approach are reported in Sidky et al.[7] and an open-source and user-friendly Python package implementing the LSS approach is available from `www.github.com/Ferg-Lab/LSS`. Here we provide a brief description of the

mathematical basis and numerical implementations underpinning each of the three steps.

2.3.1 Encoder: State-free reversible VAMPnets (SRVs)

For equilibrium systems obeying detailed balance, the transfer operator $\mathscr{T}$ propagates the probability of microstates with respect to the equilibrium distribution. $\mathscr{T}$ possesses a complete set of eigenfunctions $\{\psi_i(\mathbf{x})\}$ with real eigenvalues $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \ldots$,[5,72–74] where the first eigenvector and eigenvalue pair correspond to the equilibrium distribution and the implied timescales of higher order relaxation processes are determined by $t_i = -\tau / \ln \lambda_i$.[5] SRVs approximate these eigenvectors $\tilde{\psi}_i(\mathbf{x}) = \sum_j s_{ij}\chi_j(\mathbf{x})$ from molecular features $\mathbf{x}$ using the variational approach to conformational dynamics (VAC),[72,75] wherein deep canonical correlation analysis (DCCA)[76] is used to solve for both an optimal basis $\{\chi_j\}$ and optimal expansion coefficients s_{ij}.[5,72] Featurized molecular coordinates are used to train a twin-lobed deep neural networks to minimize a VAMP-r loss function $\mathscr{L}_{\mathrm{SRV}} = -\sum_m \lambda_m^r$,[6] where the number of slow modes to retain m is identified by a gap in the eigenvalue spectrum, and it is typical to adopt r=2. Collective couplings between the degrees of freedom in a molecular system generically give rise to a separation of time scales and typically enabling the identification of a slow subspace.[7] After training, the SRV represents an encoding E from a molecular trajectory to a m-dimensional latent space (Figure 2A). Conceptually SRVs can be considered a nonlinear generalization of time-independent components analysis (tICA)[3,77] in which the basis functions are learned from the data. SRVs are identical to the DeepTICA approach subsequently introduced by Bonati et al.[78] In applications to the PROTAC system, we employed an SRV network comprising two fully connected hidden layers of size 100, a four-dimensional latent space, and a lag time of 200 ns. Training was conducted using the Adam algorithm[79] , a batch size of 5000, a learning rate of 0.0005, and tanh activations, and 1 epoch.

2.3.2 Propagator: Mixture density networks (MDNs)

Mixture density networks (MDN) are used in the LSS framework to propagate SRV coordinates and generate low-dimensional synthetic trajectories. Rather than "memorizing" exact trajectories, MDNs use a mixture of Gaussian probability distributions to learn microscopic transition density elements $p_\tau(\boldsymbol{\psi}_{t+\tau}|\boldsymbol{\psi}_t)$ given a sufficiently large lag time time τ and a latent space $\boldsymbol{\psi}(\mathbf{x})$ spanned by the leading slow modes of $\mathscr{T}$. MDNs supplement mixture density models with deep neural networks for the specialized task of learning multimodal probability distributions.[80,81] Transition densities are represented by a linear combination of C m-dimensional Gaussian kernels $\boldsymbol{\phi}_c$,

$$p_\tau(\boldsymbol{\psi}_{t+\tau}|\boldsymbol{\psi}_t) = \sum_{c=1}^{C} \alpha_c(\boldsymbol{\psi}_t)\boldsymbol{\phi}_c(\boldsymbol{\psi}_{t+\tau}; \boldsymbol{\mu}_c(\boldsymbol{\psi}_t), \boldsymbol{\sigma}_c(\boldsymbol{\psi}_t)), \tag{1}$$

During training, the $\boldsymbol{\psi}_t$-dependent Gaussian means $\boldsymbol{\mu}_c$, variances $\boldsymbol{\sigma}_c$, and linear mixing coefficients α_c are learned from training data projected by the SRV encoder into the latent space such that the loss function $\mathscr{L}_{\text{MDN}} = -\sum_\gamma \ln p_\tau(\boldsymbol{\psi}^\gamma_{t+\tau}|\boldsymbol{\psi}^\gamma_t)$ is minimized across time-lagged training pairs γ (Figure 2B). A softmax activation is used for α_c to ensure mixing coefficients sum to unity, and $\boldsymbol{\mu}_c$ are bounded from [0, 1] by sigmoid activations. During inference, the MDN acts as a latent space propagator P iteratively samples transition densities $p_\tau(\boldsymbol{\psi}_{t+\tau}|\boldsymbol{\psi}_t)$ from some initial $\boldsymbol{\psi}_0$ to advance the system through time. Importantly, propagation is conducted entirely within the latent space meaning that the system does not require repeated decoding and encoding from configuration space in order to propagate the dynamics. This avoids the computational overhead associated with encoding and decoding at each timestep and any potential accumulation of errors from repeated decoding and re-encoding.[82,83] New trajectories are generated by a sampling transition densities approximated by transition statistics in the training data. The MDN therefore represents a microscopic generator consistent with the learned microscopic dynamics within the training data that can produce novel dynamical pathways through the latent space that are not simply carbon copies of the

training trajectories. In applications to the PROTAC system, we employed a MDN network comprising two hidden layers of size 100, swish activations, 50 Gaussian kernels, a lag time of 200 ns, and a four-dimensional latent space. We have observed that adding additional Gaussian kernels increases training time but does not offer further improvement in expressiveness or accuracy. Training was conducted for 5000 epochs using the Adam algorithm,[79] a batch size of 100,000, and a learning rate of 0.001.

2.3.3 Decoder: Conditional Wasserstein GAN (cWGAN)

The LSS decoder uses a conditional Wasserstein generative adversarial network (cWGAN) to reconstruct synthetic configurations from low-dimensional synthetic SRV coordinates.[84,85] Adversarial training is performed between a generator $G(\mathbf{z})$ that outputs molecular configurations from inputs $\mathbf{z} \sim \mathcal{P}_z(\mathbf{z})$ and a critic $C(\mathbf{x})$ that evaluates the quality of a molecular configuration $\mathbf{x}$ (Figure 2C). The networks are jointly trained to minimize a loss function based on the Wasserstein (i.e., earth mover's) distance,

$$\mathscr{L}_{\text{WGAN}} = \max_{w \in W} \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_x}[C_w(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim \mathcal{P}_z}[C_w(G(\mathbf{z}))], \tag{2}$$

where $\mathcal{P}_x(\mathbf{x})$ is the distribution over molecular configurations sampled by the MD training trajectory and $\{C_w\}_{w \in W}$ is a family of K-Lipschitz functions enforced through a gradient penalty.[84,85] To generate molecular configurations consistent with particular states in the latent space we pass $\boldsymbol{\psi}$ as a conditioning variable to G and C[86] and drive the generator with d-dimensional Gaussian noise $\mathcal{P}_z(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{1}) \in \mathbb{R}^d$. Conceptually, the cWGAN can be thought of as generating molecular configurations with particular states of the slow modes defined by a conditioning on the latent space coordinates and filling in the fast modes from the learned distribution over these modes contained within the training data. As such, the noise enables G to generate multiple molecular configurations with different states of the fast modes consistent with each conditioning latent space location that specifies the state of

the slow modes. In applications to the PROTAC system, we employed a cWGAN network comprising a generator and discriminator that each used 200 nodes per hidden layer and were trained using the Adam algorithm[79] for 400 epochs and a batch size of 5000. The noise dimension of the generator was set to 50. Gradients of the discriminator were updated five times more frequently than those of the generator.

2.3.4 Deployment

After training the three components of the LSS pipeline the encoder, propagator, and decoder are assembled back-to-back and used to generate novel synthetic trajectories at a fraction of the cost of MD. First a single initial configuration is passed through the encoder to initialize a starting coordinates in the slow latent space. The propagator is then used to efficiently generate a novel synthetic trajectory through the latent space. Importantly, since the propagator was trained to learn the transition density elements in this space that are subsequently sampled, it is not constrained to produce carbon copies of the training trajectories, but rather stochastically generates novel trajectories consistent with the statistics of the learned microscopic dynamical generator underpinning the molecular system. The latent space is intrinsically low-dimensional since it retains only the leading slow dynamical modes of the molecular system and this dynamical coarse-graining makes it both tractable to learn the transition density elements within this space during learning and extremely efficient to sample from them to propagate the dynamics during inference. Finally, the decoder operates upon the latent space trajectory produced by the propagator to generate representative molecular structures corresponding to each instance of the trajectory. One can choose to stochastically sample a single configuration corresponding to each instant or an ensemble of configurations that will retain the same slow modes but different representations of the fast modes not contained within the latent space. In applications to the PROTAC system, we employed the trained LSS pipeline to generate a 1.3 million-frame C_α trajectory at a step size of 200 ns, corresponding to 260 ms of simulations data. This constitutes $\sim$40$\times$ more

data than was contained in the Folding@Home training data, is temporally and spatially continuous single trajectory, and required only ∼4 GPU-minutes to generate on an NVIDIA GeForce RTX 2080 GPU card.

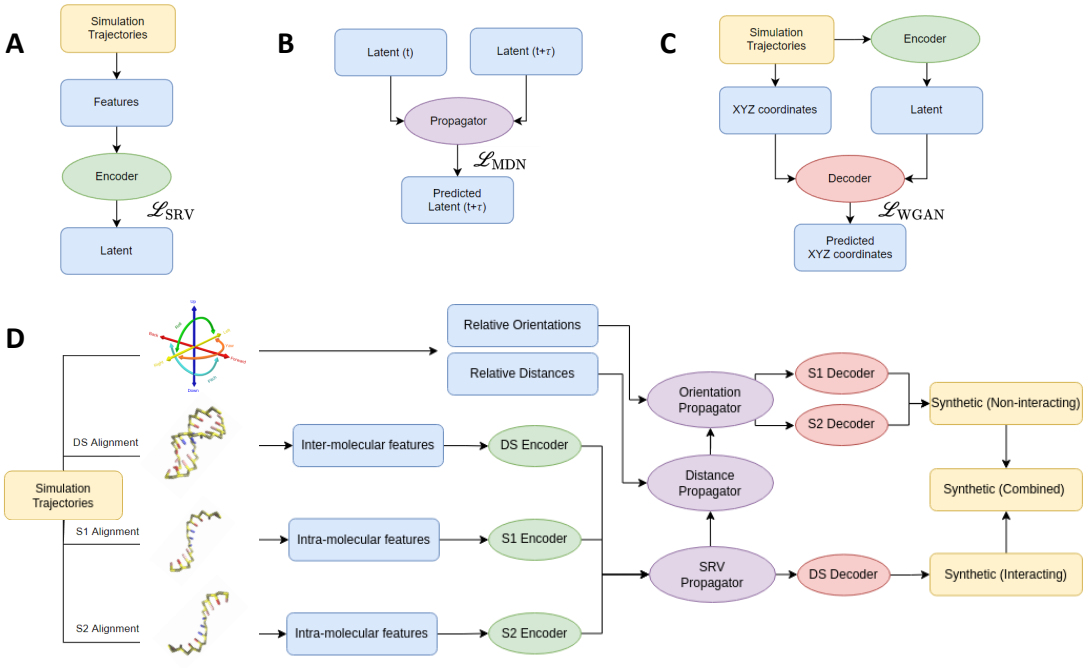

Figure 2: Schematic of LSS and Multi-LSS architecture, training, and deployment. A) The LSS SRV encoder is trained on molecular dynamics trajectories to learn a low-dimensional latent space embedding into a slow subspace spanned by the maximally autocorrelated (i.e., slowest relaxing) dynamical modes as the optimal coordinates for construction of a long-time kinetic model. B) The LSS MDN propagator is trained on time-lagged projections of training data into the latent space to predict transition densities within the learned slow latent space. C) The LSS cWGAN decoder is trained generate molecular configurations conditioned on latent coordinates. D) The Multi-LSS for a bimolecular system is trained by separating MD trajectories into four distinct components: DS – the global system comprising both molecules; S1 and S2 – the independent subsystems comprising each molecule, and DoF – the relative translational and orientational degrees of freedom defining the spatial location of S2 relative to S1. Independent encoders and decoders are trained for S1, S2, and DS. The latent spaces for these three systems are concatenated and the dynamical evolution propagated simultaneously. DoF relative distance and relative orientation propagators are trained sequentially and used to place S2 relative to S1. S1 and S2 decoding is employed when the two molecules are in a non-interacting regime and their dynamical evolution is weakly coupled. DS decoding is employed when the two molecules are interacting and their dynamical evolution is strongly coupled. A switching function is employed to distinguish which regime is active and therefore which decoding strategy to employ.

2.4 Multi-molecular latent space simulator (Multi-LSS)

Multi-molecular or loosely coupled systems have been shown to produce inefficiencies and inaccuracies when modeled by MSMs and other global kinetic models.[9,33] These issues originate from the approximately independent nature of the dynamical evolution of the constituent subsystems that leads to an exponentially growing number of states – M^N for N subsystems each containing M metastable states – that are required to describe such a system. Many of these states are physically degenerate and this state-space requires exponentially more simulation time to adequately sample and necessitates learning of an exponentially larger transition matrix. A primary outcome of this work is to address these challenges with the development of a multi-molecular generalization of the LSS (Multi-LSS) and demonstrate its application to our DNA simulations as a prototypical multi-molecular system that exhibits weak coupling between the two single strands in the dehybridized state. In the following sections, we described architectural components of the Multi-LSS to enable (i) separate encoding of each subsystem as well as the global system, (ii) serial propagators that ensure physical agreement between the global system and individual component, and (iii) a switchable decoding strategy that can reconstruct the complete structure at once in the strongly-coupled regime or each component individually in the weakly coupled regime.

2.4.1 Encoder

In a multi-molecular system, SRV encoders are susceptible to learning many degenerate modes representing closely related or identical slow processes in identical molecules. These degeneracies can obscure faster but meaningful dynamics and require much higher dimensional representations. Similar challenges have been noted in MSMs of multi-molecular systems.[9,36] To address this issue in our DNA system, we built three distinct SRVs corresponding the DS – the global system comprising both molecules – and S1 and S2 – the independent subsystems comprising each molecule individually – components of the system. Each SRV was trained using identical network hyperparameters of two hidden layers of size 100, a ReLu

activation, 50,000 batch size, 10 training epochs, and 0.001 learning rate. A lag time of 1 ns was selected based on the convergence of leading timescales and the implied timescale of the fastest relevant dynamics in each model. The two leading modes were retained for each SRV, which were combined into a six-dimensional space describing the slowest dynamics – including hybridization, fraying, and hairpin folding – of each strand and of the system as a whole. The leading two modes of the DS encoder correlated highly with hybridization and terminal fraying dynamics, respectively, while the leading modes of the S1 and S2 encoders are correlated with hybridization and hairpin folding. Although we retained an equal number of modes from each system, this is not a requirement and, in general, enough modes should be selected to adequately represent the slow system dynamics as informed by a gap in the spectrum of SRV implied time scales.

2.4.2 Propagator

In a unimolecular system, a single set of SRV modes are used for training and propagated at a specified lag time.[7] In multi-molecular systems, however, the slow dynamics of each subsystem must be preserved and propagated while maintaining correct correlations and physical constraints between each subsystems. To accomplish this for our DNA system, we concatenate the leading modes from each of the DS, S1, and S2 systems and train a unified MDN to generate synthetic trajectories through this six-dimensional latent space. Although S1 and S2 modes capture enough global information to coarsely resolve the hybridization process, we find that the intermolecular features distilled by the DS modes substantially improve this resolution and report on more subtle processes such as termini fraying, which are not captured by S1 and S2 alone. We used an architecture of two hidden layers of size 100, a batch size of 100, 100 training epochs, 0.001 learning rate, 1 ns lag time, and 50 mixture components each corresponding to a trainable mean, standard deviation, and mixture contribution which are optimized during training. In order to recover the relative orientation of S1 to S2, we also propagated the translational and rotational DoFs computed from

the training data. To ensure physical agreement with the synthetic low-dimensional space, we find that it is important to condition DoFs on past DoF coordinates as well as current SRV coordinates. Specifically, we first propagate the translations T using $p_\tau(T_{t+\tau}|T_t, \psi_{t+\tau})$ then use this information to propagate the rotations R as $p_\tau(R_{t+\tau}|R_t, T_{t+\tau}, \psi_{t+\tau})$. In this way the propagators corresponding to SRV coordinates, relative translations, and relative rotations, can be trained in series to ensure correct internal states and relative positions. Although we do not enforce detailed balance when training the MDN in this work, we have found comparable results when training on time-reversed data (Figure S1 in the Supporting Information) and note that an equilibrium data set can be augmented with time-reversed data to implicitly enforce detailed balance via a data augmentation strategy.

2.4.3 Decoder

As was done for the encoder, we train separated decoders for each of the encoded system components. For the DNA system, this means training independent decoders for DS, S1, and S2. The training configuration coordinates for each systems were aligned as described in Section 2.2. The S1 and S2 decoders were trained on all available configurations. Since the DS decoder is only used to generate configurations when the two strands are interacting we simplify its training and enhance its performance by including configurations with at least one WCF interaction (i.e., at least one pair of complementary base pairs within a 1 nm cutoff) in training. Before training, all SRV coordinates and aligned Cartesian coordinates were concatenated across trajectories and sub-sampled every 1 ns (10 frames). The same cWGAN architecture and training procedure was used for each decoder and consisted of two 100-dimensional hidden layers, 5 discriminators, a batch size of 1000, and 5000 training epochs. The latent (input) dimension was defined by total number of SRV input modes, and the output dimension was $3N$ where N is the total number of atoms in the structure (64 for DS, 32 for S1 and S2).

For each set of low-dimensional synthetic coordinates, reconstruction was performed for

the DS, S1, and S2 decoders. Although we learn a general model for the dynamics, we employ two switchable decoders to (i) capture fine-grained details of hybridized configurations or (ii) accurately place and orient dissociated strands with respect to one another. The DS decoder produced configurations of both strands directly and required no post-processing. The relative orientations of synthetic S1 and S2 were obtained from the propagated DoFs (Section 2.4.2) and the two strand were re-aligned using Procrustes alignment as detailed in Section 2.2.[52,53] These represent two independent approaches to reconstruction producing either a DS configuration and a S1+S2 configuration. Although either decoded configuration can be used, the DS approach is expected to produce higher fidelity structures when strands are interacting since it explicitly accounts for interactions between the strands in the associated state whereas the S1+S2 model treats the two strands as independent dynamical systems. Conversely, the S1+S2 model is expected to be superior when the two strands are separated in the dissociated state since each strand does now evolve as an independent subsystem and training of the DS model is compromised by the exponential explosion in number of equivalent and degenerate states of the two non-interacting strands. Empirical testing reveals the DS decoder to outperform the S1+S2 when the predicted center-of-mass between strands is less than 1 nm, and the converse to be true for distances greater than 1 nm (Figure S2). As such we employ a switching procedure such that the DS decoder is used at predicted center-of-mass distances between strands of less than 1 nm, and the S1+S2 decoder otherwise. In practice, we find the reconstruction accuracy to be relatively insensitive to the exact choice of this cutoff over the range [0.8, 1.6] nm.

2.4.4 Deployment

For the DNA system, we trained LSS models at a variety of temperatures in the range 310 K to 330 K and implicit NaCl concentrations in the range 25 mM to 400 mM. The encoder and decoder were trained only under "wild type" (WT) conditions at 320 K and 100 mM NaCl. We then used the trained encoder and decoder transferably within Multi-LSS models at other

temperatures and salt concentrations under the assumption that the state space explored by the system under these conditions is sufficiently representative of the other conditions to provide an adequate encoding and decoding under the changed conditions. As we will show, this proved to be a good approximation and validated the transferability of the encoder and decoder. To account for the altered kinetics, independent propagators were trained for each temperature and ion concentration. After training was complete, we deployed the trained LSS models under each set of conditions to generate 10×200 μs trajectories, half commencing from a hybridized duplex and half from a dissociated state, to produce $10\times$ more trajectory data than was contained in the training ensemble.

3 Results & Discussion

3.1 Application of LSS to PROTAC ternary complex

Our first goal was to train an LSS model of the PROTAC ternary complex illustrated in Figure 1 from ∼5.7 ms of short simulation trajectories of median length ∼650 ns generated by distributed computing using Folding@Home.[32] The ternary complex comprises a 149-residue VHL protein that is complexed with the 115-residue bromodomain of the SMC2 protein by the PROTAC2 molecule. The SMC2 is the degradation target of anti-cancer therapies[87] and the VHL protein is the substrate recognition domain of the multi-protein Cullin-RING E3 ubiquitin ligase (CRL) that ubiquitinates and marks for proteosomal degradation target proteins bound by VHL.[87,88] The PROTAC2 molecule is heterobifunctional, with one end comprising a binding moiety specific to VHL and the other to the SMC2 protein of interest. The PROTAC mechanism of action is to effectively modulate VHL substrate specificity by mediating its binding to proteins of interest. By recruiting VHL into a complex with SMC2, the CRL is brought into proximity of SMC2 to ubiquitinate and selectively mark it for proteosomal destruction.

Prior work by Dixon et al.[24] suggests that the thermodynamics of the bound ternary com-

plex mediated by the PROTAC molecule are an insufficient correlate of the observed degradation efficiency for the three PROTAC molecules PROTAC1, PROTAC2, and ACBI1. Of these three molecules, distinct binding poses of SMC2 relative to VHL were associated with ACBI1 and PROTAC1, and ACBI1 promoted higher degradation activity than PROTAC1. The ternary complex induced by PROTAC2 was hypothesized to oscillate between the two binding poses and underpin the observed degradation activity intermediate to PROTAC1 and ACBI1. Taken together, this led Dixon et al. to propose a dynamical basis for PROTAC activity based on the relative stability and net residence time within a binding pose of SMC2 relative to VHL favorable to CRL ubiquitination.[24,87] This ensemble is challenging to study experimentally given that its constituent states may not be favored by crystallization,[89,90] but it is also non-trivial to quantify and sample the ensemble computationally without appropriate collective variables.

These findings motivated the two objectives for the present work. First, to demonstrate training of LSS models over short trajectories generated by distributed computing and use of the trained model to efficiently produce long spatially and temporally continuous molecular trajectories. For large molecular systems such as the PROTAC ternary complex, these single, long trajectories would be exceedingly expensive – if not impracticable – to produce by direct molecular simulation, but are very valuable in exposing the dynamical transitions and putative transition states between metastable minima and enabling dense sampling of transition pathways to achieve good statistical power in kinetic estimates of rate constants and dwell times. Second, we seek to determine if the LSS pipeline, and in particular the SRV encoder, can identify and parameterize the important configurational motions characterizing the transition between the two metastable VHL/SMC2 binding poses mediated by PROTAC2. We hypothesize that if this is a sufficiently slow transition, that the SRV should be capable of learning this dynamical motion in a completely unsupervised manner and discovering a collective variable (CV) parameterizing this transition as a component of the learned latent space. The trained LSS model can furnish an efficient dynamic simulator

to interrogate and sample these transitions to both expose mechanistic understanding of this conformational change and furnish a learned slow CV as a good discriminant of the two poses and putative metric for rational design of novel PROTAC molecules by computational screening.

SRV encoder learns slow collective variables to explain differential PROTAC degradation efficacy

We trained a LSS model for the VHL-PROTAC2-SMC2 ternary complex over 5.7 ms of discontinuous trajectories generated by distributed computing using Folding@Home.[32] We first interrogate the slow modes learned by the trained SRV encoder that define the slow latent space and correlate these with particular conformational motions of the ternary complex. Four leading slow modes were resolved by the SRV encoder resulting in a 4D latent space $\boldsymbol{\psi} = \{\psi_0, \psi_1, \psi_2, \psi_3\}$. We illustrate in Figure 3 a projection of the aggregated 5.7 ms of training trajectories into the latent space by passing each configuration through the trained SRV encoder. Corresponding free energy landscapes are presented in Figure 4. The latent space possesses five metastable basins containing particular metastable macrostates of the ternary complex. In some cases the discontinuous training trajectories densely sample the intermediate configurations linking these basins, and in other cases the transition pathway is more sparsely sampled. A key benefit of the LSS paradigm is to patch together the dynamical information contained within these discontinuous trajectories into a single kinetic model capable of generating long temporally and spatially continuous trajectories that can inexpensively generate transitions between these states. Of course, transitions that are sampled more densely in the training data are better parameterized in the LSS model, whereas the model is forced to interpolate intermediate configurations between metastable states for which transitions are less well represented in the training data. We return to this point below and discuss means to adaptively improve sampling undersampled transition states in the Conclusions.

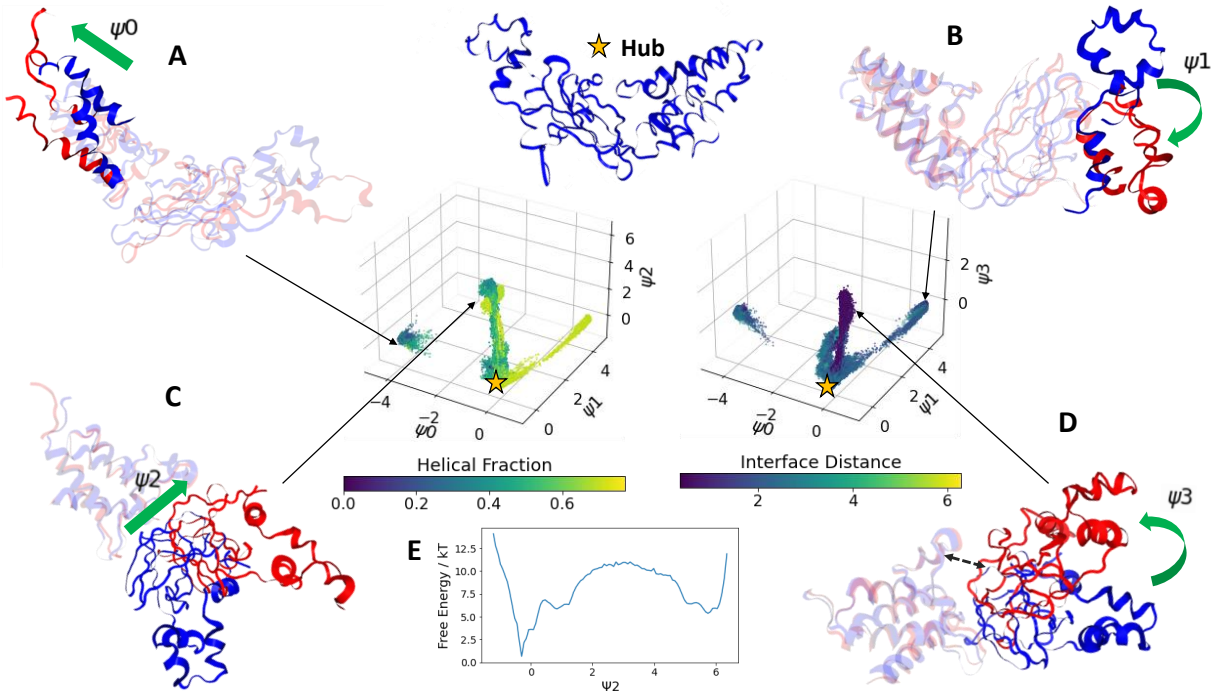

Figure 3: Analysis of slow latent space of the VHL-PROTAC2-SMC2 ternary complex. A-D) Projecting the 5.7 ms of Folding@Home training data into the 4D latent space exposes a simple "starburst" topology in which the metastable states are located at the termini of spikes emerging from a central hub located at the origin. The embedded points are colored by the helical fraction of the N-terminus SMC2 helix (left) and the "interface distance" (black arrow in D) measuring the linear displacement between the top of the SMC2 helix and the tip of the VHL tongue (right). Representative configurations at various locations in the latent space (red) are superposed on a representative configuration from the central hub (blue) to illustrate the structural changes associated with excursions along each of the learned slow modes $\{\psi_0, \psi_1, \psi_2, \psi_3\}$ spanning the axes of the latent space. Structural variations of interest relative to the hub are shown by opaque coloring, and the remaining structure is show transparently. Green arrows are labelled by the correspond ψ coordinate and highlight the following transitions associated with each metastable state: A) folding/unfolding of the two SMC2 terminal helices, B) stacking/unstacking of two helices near the C-terminus of VHL, C) hinged opening/closing of the VHL-SMC2 interface, D) screw-like rotation of the VHL-SMC2 interface. E) Free energy surface of the training data projected into ψ_2.

The projection of the training data into the latent space defines a manifold with a relatively simple "starburst" topology in which the metastable states are located at the termini of spikes emerging from a central hub located at the origin $(0, 0, 0, 0)$. This structure indicates that transitions between metastable states tend to proceed through the central hub rather than via direct interconversions. To gain intuition into the conformational motions associ-

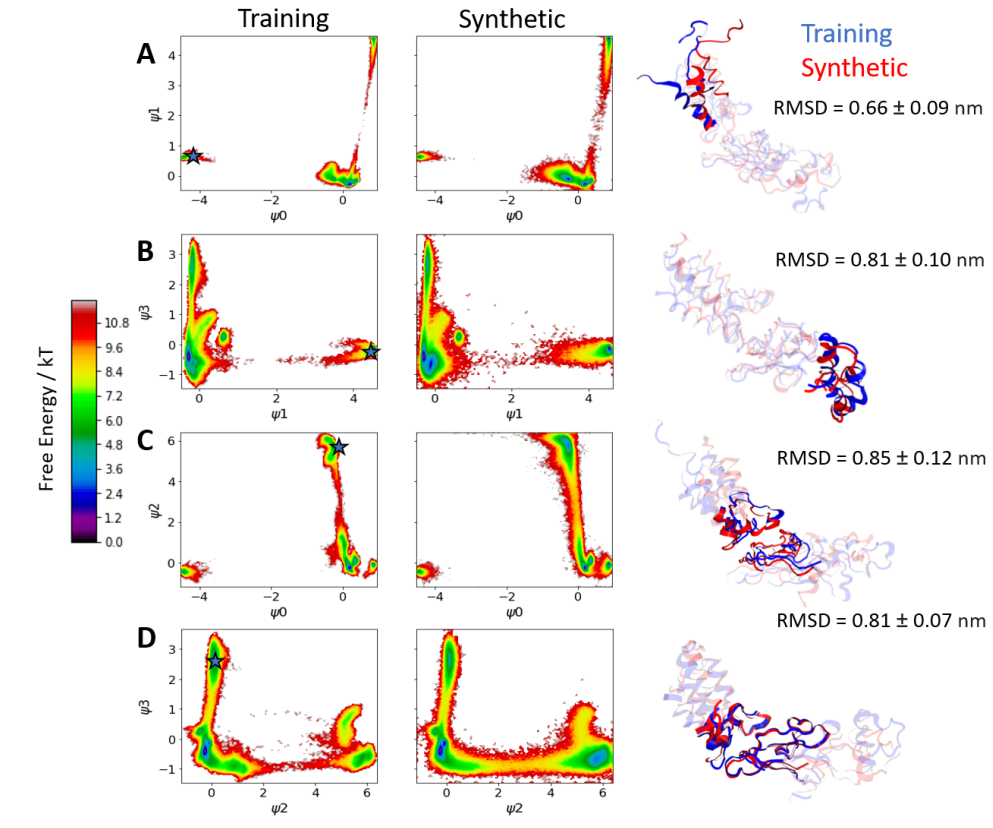

Figure 4: Thermodynamic and structural comparison between the 5.7 ms of discontinuous Folding@Home training trajectories and the continuous 260 ms synthetic LSS trajectory containing the same number of frames. A-D) Free energy surfaces are estimated as $\beta F(\boldsymbol{\psi}) = -\ln P(\boldsymbol{\psi}) + C$, where $P(\boldsymbol{\psi})$ is the empirical probability distribution of the training data projected into the 4D slow latent space, $\beta = (k_B T)^{-1}$ is the reciprocal temperature, and C is an arbitrary additive constant reflecting our ignorance of the absolute free energy scale. For visual clarity, we present 2D projections of the 4D free energy landscape constructed by marginalizing over the omitted dimensions. Since only free energy differences are meaningful, we adjust the arbitrary additive constant in each plot such that the minimum free energy is zero. The training and synthetic free energy landscapes are in good agreement. Representative structures from the hold-out training (blue) and synthetic (red) trajectories are extracted from the same location in the latent space indicated by the star, translationally and rotationally superposed, and the C_α RMSD reported. Opaque residues highlight the regions of interest within the ternary complex associated with structural changes relative to the hub configuration at the origin of the latent space (cf. Figure 3): A) unfolding of the two SMC2 terminal helices, B) stacking of two helices near the C-terminus of VHL, C) hinged opening of the VHL-SMC2 interface, D) left-handed screw-like rotation of the VHL-SMC2 interface. The synthetic configurations faithfully reproduce the structural changes exhibited in the training data.

ated with excursions along each of the four slow modes $\boldsymbol{\psi} = \{\psi_0, \psi_1, \psi_2, \psi_3\}$ parameterizing the axes of the slow latent space, we color the embeddings according to candidate physical variables. We color the left plot in Figure 3 presenting the $\{\psi_0, \psi_1, \psi_2\}$ projection by the helical fraction of the N-terminus SMC2 helix, and the right plot presenting the $\{\psi_0, \psi_1, \psi_3\}$ projection by the "interface distance" measuring the linear displacement between the top of the SMC2 helix and the tip of the VHL tongue. We also visualize a number of representative configurations at selected locations in the latent space as red ribbon diagrams, and overlay these with the hub configuration rendered in blue. We use opaque shading to highlight the distinct structural characteristics of each metastable state relative to the hub and green arrows to highlight the structural difference between the two configurations associated with excursions along a particular dimension of the latent space and therefore the physical CVs associated with each learned slow mode.

Commencing from the central hub configuration, excursions along ψ_0 to the metastable state located at approximately $(-4, 1, 0, 0)$ correspond to partial unfolding of helices at the N-terminus and C-terminus of SMC2 (Figure 3A). We note that there is a discontinuity along ψ_0 reflecting the absence of transitions along this mode in the training data, and we discuss below the impact of this gap on the kinetics. Excursions along ψ_1 from the central hub to the metastable basin at approximately $(1, 4, 0, 0)$ correspond to stacking of helices at the C-terminus of VHL (Figure 3B). Accordingly, the first two learned slow modes correspond to conformational changes within the SMC2 and VHL proteins distal from their binding interface mediated by the PROTAC2 molecule. The VHL helix stacking associated with the second mode may be relevant to recruitment/interaction with the other members of the CRL complex, but are likely to be suppressed once VHL is complexed with Elongin B, Elongin C, and Cul2 within the CRL (Figure 1). Additional simulations of VHL within the CRL would be necessary to resolve this issue.

The conformational changes associated with excursions along ψ_2 and ψ_3 correspond to global processes centered on the binding interface that change the relative orientations of

SMC2 and VHL. Excursions along ψ_2 to the basin at approximately $(0, 0, 5, 0)$ are correlated with hinged opening of the SMC2-VHL interface (Figure 3C), driven by the loss of intermolecular contacts between SMC2 and the VHL tongue. Excursions along ψ_3 to the basin at approximately $(0, 0, 0, 3)$ are correlated with a left-handed screw-like motion of VHL relative to SMC2 (Figure 3D). The hinge-like collective motions along the learned slow modes ψ_2 is valuable understanding and rationalizing the SMC2 degradation efficacy of PROTAC2. Comparing Figure 3C to Figure 7e in Ref.,[24] we observe that ψ_2 is well-suited to describe the transition between two distinct poses of VHL and SMC2 when linked by PROTAC2. In particular, the closed-hinge configuration at the hub state (Figure 3C, blue) is similar to the pose induced by the ACBI1 linker with higher degradation efficiency, and the open-hinge configuration is similar to that induced by the PROTAC1 linker with lower degradation efficiency (Figure 3C, red). Dixon et al. hypothesized that the intermediate degradation efficiency of PROTAC2 relative to ACBI1 and PROTAC1 resulted from its dynamic oscillation between both of these binding poses, one of which (closed-hinge, ACBI1) is more favorable in promoting ubiquitination in the CRL than the other (open-hinge, PROTAC1).[24]

The 1D free energy landscape in ψ_2 provides an estimate of the relative stability of these two binding poses (Figure 3E). Sampling multiple ligase-target binding poses permits the design of PROTAC molecules according to these different poses, thus amplifying the opportunity of a successful design. Under Dixon et al.'s hypothesis, shifting the population to the closed-hinge configuration should improve degradation efficiency. As such, this landscape presents a means to predict the degradation efficiency of novel PROTAC molecules from molecular simulation data by estimating the relative stability of the desired closed-hinge binding pose. This presents a putative objective function and optimization metric for the computational design and evaluation of novel PROTAC molecules. Moreover, the ψ_2 CV learned by the SRV could be used as an order parameters for enhanced sampling calculations to accelerate sampling of this rare transition to reduce the computational cost of the all-atom molecular simulations necessary to achieve converged thermodynamic stability

estimates.[78,91–93]

In sum, one of the slow modes learned from the training data by the SRV in an unsupervised manner and without any prior physical knowledge characterizes the structural changes associated with dynamical motions between open-hinge and closed-hinge VHL-SMC2 binding poses. The relative population of these poses has been conjectured to correlate with PROTAC degradation efficiency, and the free energy landscape in this learned CV offers a quantitative metric for PROTAC evaluation and design. Finally, the learned CV itself can be implemented within enhanced sampling calculations to accelerate convergence of molecular dynamics simulations and enable a high-throughput virtual screening (HTVS) campaign for *in silico* screening of PROTAC candidates.

LSS generates realistic ultra-long simulation of PROTACs

We now test the thermodynamic and kinetic consistency of the trained LSS model and use it to efficiently generate ultra-long continuous simulation trajectories to predict, expose, and densely sample conformational mechanistic changes. The LSS model was trained over 5.7 ms of discontinuous Folding@Home trajectories produced at a cost of ∼7000 GPU-years, but the trained model can generate a 260 ms spatially and temporally continuous trajectory containing the same number of frames in less than 4 GPU-minutes. Short segments of the LSS trajectory showing the conformational transitions associated with each of the four dimensions of the latent space are provided as Movies S1-4 in the Supporting Information. Having produced this long synthetic trajectory, we now assess its thermodynamic, structural, and kinetic consistency with the training data.

Thermodynamically, we compare in Figure 4 the free energy landscapes over the slow latent space constructed from the 5.7 ms Folding@Home training data and the 260 ms synthetic LSS trajectory. For visual clarity, we present 2D projections of the full 4D free energy landscape. Despite being initiated from a single seed, the synthetic trajectory fully explores all of the metastable states of the system, and the location and relative stability of

these states are in good agreement between the training and synthetic data. As anticipated from the topology of the latent space, transitions between the metastable minima residing at the tips of the "spikes" in each latent space dimension always proceed through the hub state located at the origin. The length of the LSS trajectory means that thousands of transitions in and out of the hub state are observed, and each metastable state is visited a minimum of 75 times. The synthetic data exhibits slightly denser sampling of the interstitial regions between the metastable basins. The transition pathways by their very nature tend to lie in high-free energy regions of configurational phase space and are therefore only transiently and sparsely sampled in the Folding@Home training data. It is not surprising, therefore, that the free energy predictions of the LSS model in these regions are in less good agreement than in the metastable free energy minima. Furthermore, although the Folding@Home trajectories had a median length of ∼650 ns, these appear to have been sufficiently long to sample all relevant dynamical transitions between metastable states and furnish a fully-connected LSS model capable of visiting all relevant regions of the latent space.

Structurally, we assessed the decoder's ability to properly reconstruct molecular structures from the latent space. We present next to each pair of free energy landscapes in Figure 4 representative configurations harvested from the training and synthetic data from various metastable minima in the latent space free energy surfaces. We encoded and decoded the 60,000 frames within the 5% hold-out test partition not included in the LSS training pipeline and computed the mean C_α root mean squared deviations (RMSD) between the true and reconstructed configurations. This data set contained at least 780 frames in each of the metastable minima discussed above. We observe very good agreement between the training and synthetic structures in all metastable minima, achieving C_α root mean squared deviations (RMSD) of (0.85 ± 0.12) nm or better. Visual inspection confirms that the expected conformational changes associated with these metastable minima (cf. Figure 3) are correctly predicted by the synthetic LSS structures. Overall, low RMSD values show that the LSS decoder produces a distribution of predicted molecular structures in good agreement with

the distribution in the training data, and local structural agreement as a function of $\boldsymbol{\psi}$ shows the distribution is properly conditioned by the latent space coordinates. Conceptually, the $\boldsymbol{\psi}$ conditioning informs the decoder as to the status of the four slow degrees of freedom identified by the SRV and encoded in the latent space, and the decoder then generates an ensemble of molecular configurations with different fast degrees of freedom that are annealed to the state of the slow collective variables and correctly match the distribution observed in the training data.

Kinetically, we compare the LSS model with the training data in two ways. First, we computed the autocorrelation times associated with the four slow modes $\{\psi_0, \psi_1, \psi_2, \psi_3\}$ calculated from the 5.7 ns of discontinuous Folding@Home trajectories and compare these with the implied timescales learned by the trained SRV encoder (Table 1). The good agreement between these values indicates that the SRV encoder is faithfully learning the dynamical relaxations associated with these leading slow modes. Second, we computed the slow mode autocorrelation times from the synthetic LSS trajectory. These timescales are in similar agreement with those computed from the training data, with the exception of the leading slow timescale associated with ψ_0, which was accelerated by almost an order of magnitude in the synthetic data. We attribute this to the fact that no transitions along ψ_0 were observed in the training data, and therefore we do not possess an accurate ground-truth estimate for this timescale. Interestingly, despite there being no transitions in the training data, the MDN propagator learns to make transitions in ψ_0 to sample to and fro between the central hub configuration at the origin of the latent space and the metastable state with unfolded SMC2 helices at approximately $(-4, 1, 0, 0)$. Empirically, we observe that during training the propagator initially transitions between all states very rapidly and first optimizes thermodynamic agreement, it then gradually slows the kinetics down to approach those of the training data. The propagator is, of course, not capable of learning the kinetics for unsampled transitions, and the learned LSS time scales are therefore an unconverged estimate for a ground truth time scale to which we do not have access in the training data. We expect

that if we were to train the MDN propagator for a substantially longer number of epochs, the LSS may cease hopping across ψ_0 entirely to reflect that this transition does not occur in the training data. The timescale agreement in the higher-order modes is much better and we attribute this agreement to superior sampling of transitions in $\{\psi_1, \psi_2, \psi_3\}$. Somewhat fortuitously, the synthetic trajectory timescale of the ψ_2 mode – the mode associated with VHL-SMC2 interface hinging and the most relevant mode for PROTAC engineering – is in agreement to within error of that estimated from the training data. Further work is warranted to better understand the effect of median trajectory length and undersampling of transition pathways in training LSS models on short, discontinuous training trajectories, but this application has demonstrated that it is possible to train a converged LSS model with excellent thermodynamic and structural accuracy, and good kinetic accuracy except in the case of very sparse sampling of particular conformational changes.

Table 1: Comparison of the timescales of the learned slow dynamical modes calculated from (i) the encoder eigenvalues, (ii) the Folding@Home training data autocorrelation times, and (iii) the LSS synthetic data autocorrelation times.

Leading kinetic Timescales				
Leading Modes	ψ_0	ψ_1	ψ_2	ψ_3
SRV Leading Timescales	384	253	73.5	17.9
Training Autocorrelations	332 ± 50.0	280 ± 67.5	67.0 ± 16.8	12.7 ± 0.7
Synthetic Autocorrelations	58.5 ± 18.7	187 ± 24.8	79.2 ± 3.4	8.4 ± 0.7

3.2 Application of Multi-LSS to DNA system

Our second goal was to train a Multi-LSS model for a DNA system that reversibly folds into duplex and hairpin states under equilibrium conditions. The duplex state tends to be strongly favored over the hairpin state due to more stabilizing WCF base pairings and base stacking interactions. This equilibrium towards the duplex states motivates the use of hairpins as "molecular fuel" in various DNA nanotechnology applications,[42,43] and their relative stabilities can be modulated by changing environmental conditions,[43] chemically modifying base pairs,[94] or selectively adding mismatches to destabilize the hybridized state.[95–97] In

this work, we employ the latter technique, adding three mismatched T-T base-pairs to an otherwise G-C rich, self-complementary DNA duplex. The resulting system consists of two identical DNA strands 5′-GCGGTTTCCGC-3′ capable of forming up to eight complementary WCF pairs in the duplex state and four pairs in the hairpin state. We model this system using the computationally efficient coarse grained 3SPN.2 model[58] that enables us to relatively easily conduct 200 μs of simulations. Nevertheless, the hybridization/dehybridization dynamics are sufficiently slow that we only sample $\sim$20 of each event within these training data. This motivates the construction of efficient Multi-LSS simulators to learn the microscopic dynamics of these processes over relatively modest numbers of dynamical events, and then use the trained model to efficiently and densely sample these rare events to gain improved statistical resolution of kinetic observables.

Multi-LSS trajectories preserve DNA structures and thermodynamics

We first trained a Multi-LSS on 10×20 μs $= 200$ μs training trajectories collected under "wild type" (WT) conditions of 320 K and 100 mM NaCl concentration. We then employed the trained model to produce 10×200 μs $= 2$ ms spatially and temporally continuous synthetic trajectories. We first test the thermodynamic and structural consistency of the synthetic trajectories with the training data. In Figure 5 we present a comparison of the free energy surfaces collected over the 200 μs training data and 2 ms synthetic data. We project these into a consistent basis defined by the two leading TICA modes (TICs) computed over a translationally and rotationally invariant featurization comprising the 231 intra-molecular and inter-molecular distances as a more visually interpretable low-dimensional embedding of the 6D SRV latent space.[3,98,99] The training and synthetic data both occupy the same phase space volume and define a bimodal free energy surface comprising a deep global minimum at $(-1.3, 0)$ containing the duplex state and a broader, shallower local minimum centered at $(0.6, 0)$ containing the dehybridized state. The dehybridized basin exhibits two lobes at $(0.6, 1)$ and $(0.6, -1)$ corresponding to the single-strand hairpin configurations. These

lobes are slightly less pronounced in the synthetic data and – as was observed for the ternary complex – the transition pathway between the two metastable states are slightly oversampled by the Multi-LSS. As we discuss below in our analysis of the kinetic predictions of the model, this leads to a mild ∼10% acceleration of the hairpin transitions in the trained LSS model relative to the training data. Nevertheless, the landscapes are in quantitative agreement to within better than ∼1 kT. Representative structures visualized at particular locations over the embedding are also in excellent agreement between the training and synthetic data, with the synthetic model generating physically realistic configurations with properly formed intra-molecular and inter-molecular bond lengths, angles, and energies. In Figure S3 we present a comparison of the distribution of WCF bond lengths, hairpin bond lengths, and backbone dihedral angles and observe excellent agreement between the training and synthetic data.

We further probe the thermodynamic and structural properties of the two models by analyzing the distribution of hybridized vs. dehybridized states for the pair of strands and hairpin vs. coil configurations for each single strand. In Figure 6A we color the TICA embeddings of the training and synthetic data according to three physical variables: the inter-strand center-of-mass distance (ISD), the end-to-end distance of the first strand (EE1), and the end-to-end distance of the second strand (EE2). The heatmaps visually confirm that TIC1 is strongly correlated with duplex formation and TIC2 with hairpin formation, and motivate the construction of Figure 6B, in which we project all training and synthetic data into ISD, EE1, and EE2. Distributions exposed by these histograms are very similar between training and synthetic, particularly for the ISD distributions characterizing the relative proportion of hybridized and dehybridized states. The EE distributions are also in good agreement and place the probability maxima in the correct locations, but the shapes of the distributions are slightly different and the synthetic data possess slightly longer low-EE tails. The longer tails appear to be correlated with the generation of slightly more structurally diverse hairpin states generated by the Multi-LSS. In Figure 6C we integrate over the population below a distance cutoff of 1.5 nm for ISD, EE1 and EE2 to determine the fraction of duplex config-

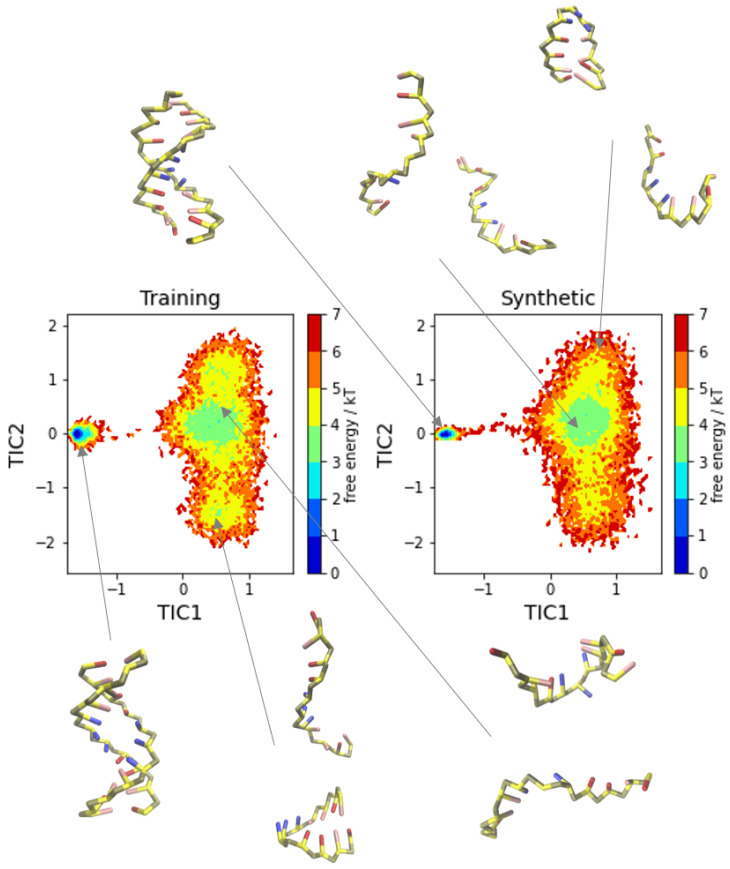

Figure 5: Thermodynamic of free energy landscapes computed by projecting the 200 μs training and 2 ms Multi-LSS synthetic trajectories for the DNA system projected into a 2D TICA embedding. The training and synthetic data explore the same volume of phase space and the relative free energy values are in quantitative agreement within $\sim$1 kT . The landscaped define a bimodal topography with the deep global minimum on the left containing the duplex state and the broader local minimum on the right containing the dehybridized and hairpin states.

urations, S1 hairpins, and S2 hairpins, respectively. We calculate these fractions over each trajectory independently and show the standard deviation between trajectories as error bars. Because each synthetic trajectory contains 10$\times$ more data than a training trajectory, the Multi-LSS allows us to substantially reduce our statistical uncertainties in reporting these values. With respect to duplex formation, we report a melted fraction of (0.29$\pm$0.05) in the synthetic data compared to (0.26$\pm$0.16) from training data. The estimated fraction of folded hairpins in the synthetic data is consistent between S1 and S2 at (0.28$\pm$0.01) and in

agreement with the value of (0.29±0.05) in the training data.

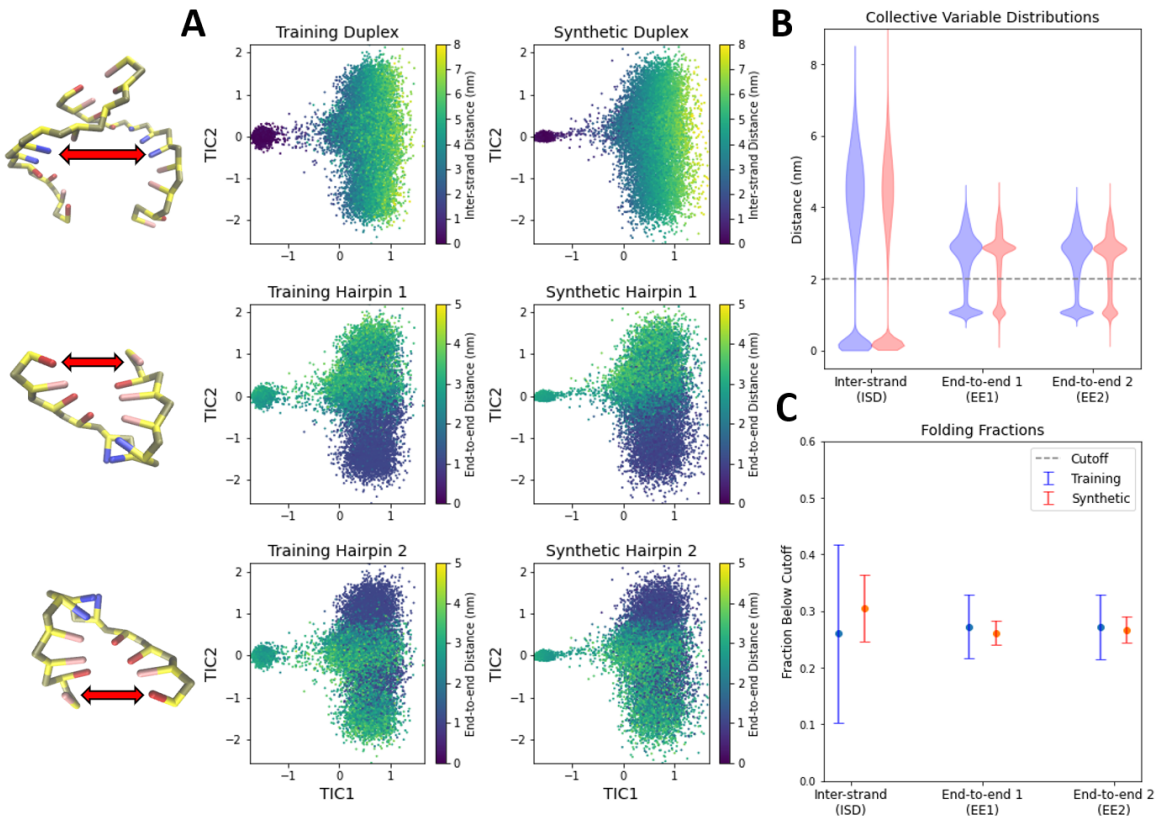

Figure 6: Structural comparison of the 200 μs training and 2 ms Multi-LSS synthetic trajectories for the DNA system. A) The 2D TICA embedding colored by three physical CVs corresponding to the inter-strand center-of-mass distance (ISD), the end-to-end distance of the first strand (EE1), and the end-to-end distance of the second strand (EE2). The embeddings are consistent between training and synthetic trajectories. B) Violin plots of the training and synthetic distributions along these three physical coordinates to expose the distributions of hybridized/dehybridized states for the pair of strands (ISD) and hairpin/coil configurations for each single strand (EE1, EE2). C) Fraction of each CV that in a hybridized or hairpin state as determined by a distance cutoff of 1.5 nm. The 10× longer synthetic trajectories permit an approximately $\sqrt{10}\times$ reduction in the estimated statistical uncertainties.

Multi-LSS trajectories preserve DNA hybridization and hairpin folding kinetics

After validating the thermodynamics of synthetic trajectories, we next evaluated if the Multi-LSS preserves the correct kinetics of the system. While a naive MDN propagator could

reproduce an equilibrium thermodynamic distribution by generating correctly distributed configurations in an erroneous order, maintaining correct kinetics represents a more substantial challenge and validation of a properly trained Multi-LSS pipeline. In Figure 7A and B we show segments of training and synthetic trajectories as a function of ISD and EE1 distances. Hybridized duplex and folded hairpin states are marked by sustained decreases in ISD and EE1 values, respectively, and we observe qualitative similarities between the training and synthetic in terms of both the magnitude of the fluctuations and and the dwell times in each state and frequency of the transitions. In Figure 7C, we quantify kinetics of ISD, EE1, and EE2 by computing their autocorrelation curves within the training and synthetic trajectories as a function of lag time. These autocorrelation times furnish implied timescales associated with the dynamical processes associated with these physical order parameters, respectively, duplex hybridization/diffusion and hairpin folding of each strand. We find the hybridization timescale to be about an order of magnitude slower than individual hairpin folding timescales. We observe excellent agreement of the implied timescale curves between the training and synthetic data, demonstrating that the Multi-LSS has faithfully learned the dynamics associated with duplex (de)hybridization and hairpin (un)folding. Again, we also note the greatly reduced statistical uncertainties associated with the synthetic data compared to the training data due to the 10× increased data volume.

In Figure 8 we present a comparison of hybridization events extracted from the training and synthetic trajectories. Hybridization can proceed by multiple different pathways depending on the particular orientation of colliding strands. The synthetic trajectory generates physically realistic hybridization events that are not simply learned copies of the events within the training data, demonstrating that the trained Multi-LSS model has learned the underlying microscopic dynamics of the molecular system within the MDN propagator, and can sample from these dynamics to produce novel molecular events.

It is a key property of the LSS paradigm to learn the microscopic generators of the molecular dynamics from relatively modest trajectory data and numbers of rare dynamical

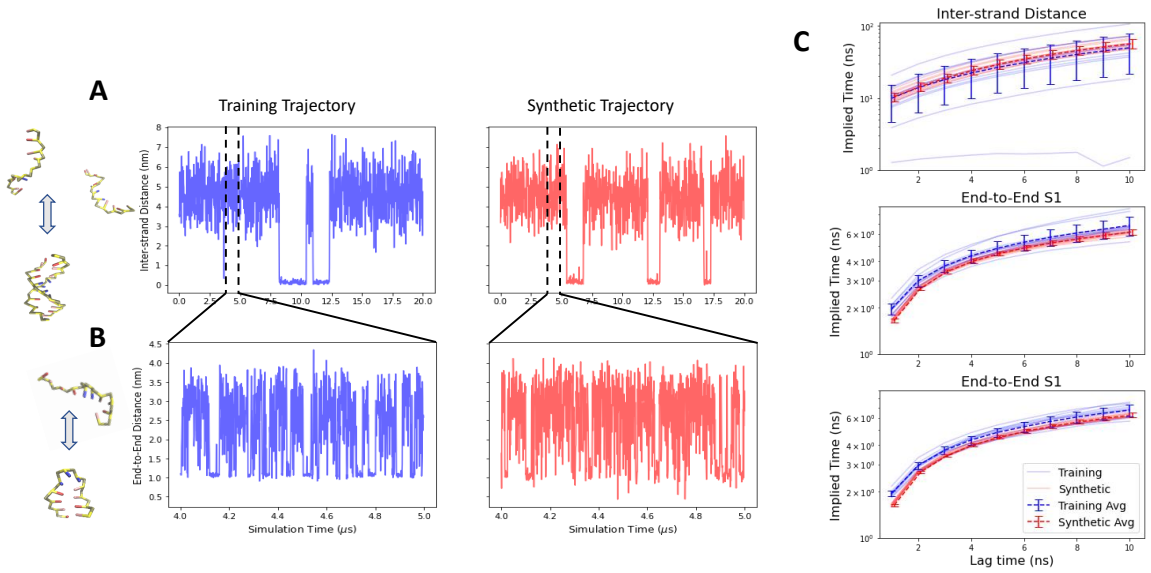

Figure 7: Kinetic comparison of the 200 μs training and 2 ms Multi-LSS synthetic trajectories for the DNA system. A) Inter-strand center-of-mass distance (ISD) for representative 20 μs regions of training and synthetic trajectories. Hybridized regions are shown by sharp reductions in ISD and suppression of fluctuations. B) End-to-end distance (EE1) of the first strand for representative 1 μs regions of training and synthetic trajectories. Hairpin states correspond to low EE1 values. C) Implied timescale plots for reporting the autocorrelation time of the physical variables ISD, EE1, and EE2 as a function of lag time. The light solid lines represent ten equal length trajectories of length 20 μs for the training data and 200 μs for the synthetic data. Heavy dashed lines represent means equipped with associated standard errors computed over these ten replicates. Note the higher variance and greater prevalence of non-representative, outlier trajectories associated with the ISD implied timescales due to the slow hybridization/dehybridization dynamics.

events – in this case, ∼20 (de)hybridization events – and then enable the generation of novel synthetic trajectories at orders of magnitude lower cost. We chose to generate only 2 ms of synthetic data for the purposes of comparison, but the exceedingly low 4 GPU-minute cost to do so means that it is possible to generate vastly more data and drive the statistical uncertainties in our dynamical observables towards zero. There are, of course, systematic uncertainties that are not captured within our standard error estimates, and we provide some discussion and commentary on these factors in the Conclusions.

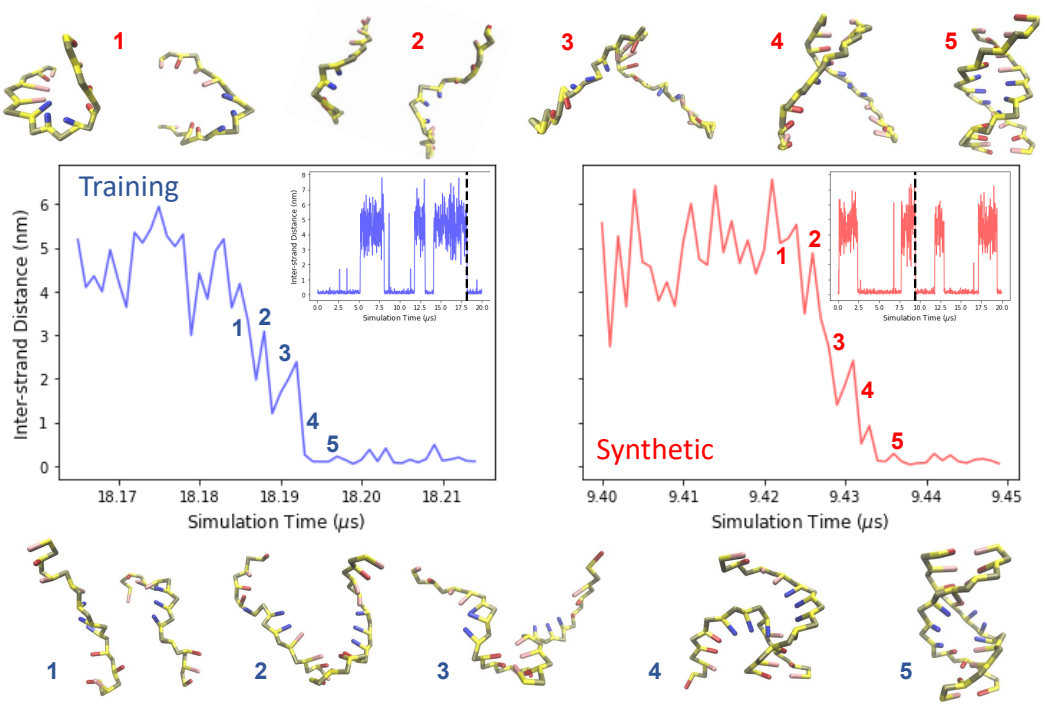

Figure 8: Visualization of representative DNA hybridization events extracted from the training and synthetic trajectories. We select as an example for comparison a hybridization event that adopts frayed configurations for several ns before transitioning to a fully hybridized state (inset) since these events proved to be particularly challenging for the Multi-LSS to correctly simulate due to the relative paucity of training data and the importance in selecting an appropriate cutoff in switching between the DS and S1+S2 decoders. The synthetic trajectory produces physically realistic hybridization events that are not just carbon copies of those present in the training data and are indistinguishable by eye from real simulated events.

To understand how synthetic trajectories transition between various metastable states, we constructed MSMs using an SRV embedding trained on all inter-molecular and intra-molecular features in the training data. We used a consistent k-means micro-state clustering procedure and PCCA+[100] macro-state state assignments for both training and synthetic data. The limitations of using MSMs to fully describe this system were discussed previously[9,36] and they do not generate temporally continuous synthetic trajectories like an LSS, but they do provide a robust method to compare the macrostate jump dynamics between the training and synthetic trajectories. Chapman-Kolmogorov (CK) tests and implied time scale plots for the two MSMs are reported in Figures S4-S7. For each MSM we computed

the stationary distributions of seven metastable macrostates and mean first passage times (MFPT) between each state. We observe excellent agreement between the macrostate free energy assignments between the training and synthetic data, with a mean difference of just 0.9 kT and a maximum 1.8 kT discrepancy over the seven macrostates (Figure 9). We also see good agreement in the MFPTs between macrostates with low wait times out of frayed and hairpins states and much longer wait times for (de)hybridization transitions (Figure 9B). While hairpin folding timescales are within error of the training data, (de)hybridization processes are slightly accelerated, likely because some transition region configurations that are oversampled in the synthetic data are clustered into the hybridized state. This is also consistent with a mild $\sim$2$\times$ acceleration of the leading implied timescales of the MSM built from synthetic data relative to that constructed over the simulation data (Figures S6-S7), although this may also be due to the elimination of fast dynamical modes during encoding and/or the reduced temporal resolution of the synthetic trajectory relative to the simulation data. Overall, however, the hierarchy of MFPTs from each state is preserved (Figure S8), indicating that kinetic behavior is well reproduced by the synthetic Multi-LSS trajectories.

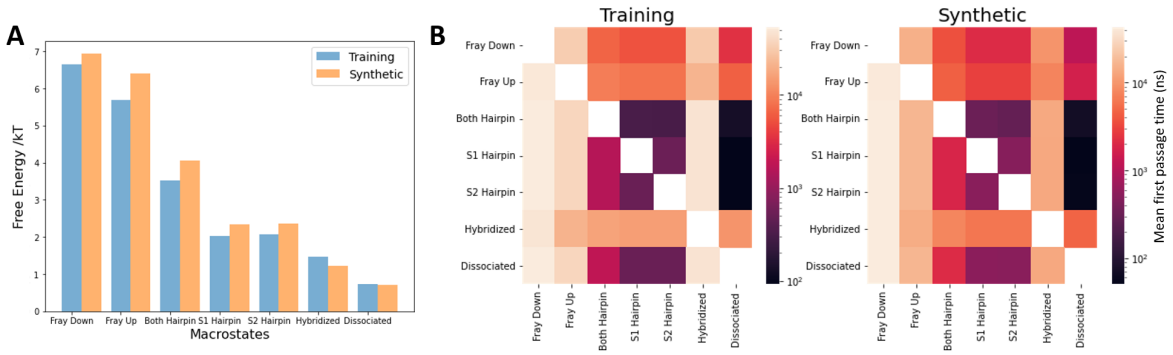

Figure 9: Markov state model (MSM) comparison of the 200 μs training and 2 ms Multi-LSS synthetic trajectories for the DNA system. A) Stationary probabilities for each MSM macrostate computed from the training and synthetic SRV-MSMs. B) Mean first passage times between each pair of SRV-MSM macrostates.

Multi-LSS transfer learning to different temperatures and salt concentrations

After performing structural, thermodynamic, and kinetic validation of the Multi-LSS pipeline on a system under the "wild type" (WT) conditions of 320 K and 100 mM NaCl concentration, we used a transfer learning approach to extend the framework to new temperatures and ionic strengths. Previously it has been shown that the same time-lagged, low-dimensional space can be used to encode or bias a related system.[93,101] We adopt a similar approach by using the same SRV encoders and cWGAN decoders trained on the WT data, but retrain the MDN propagators on new trajectory data. The underlying principle for transferability is that the state space explored by the system under the WT conditions is sufficiently representative of the perturbed conditions that the learned SRV slow latent space remains a good representation of the slow dynamics of the perturbed systems and that no substantially new molecular configurations emerge that would require retraining of the decoder. The only update to the model is to update the MDN propagator to learn the new dynamics over the same latent space. This approach can potentially reduce the volume of training data and training time required to parameterize new systems in related Multi-LSS models. The ability to reuse components of a pre-trained LSS model is particularly valuable for DNA systems where simulation data can suffer from from a strong imbalance in hybridized or dissociated data when the system is pushed away from the melting temperature, and the thermodynamics and kinetics are highly sensitive to changes in temperature and ion concentration.[102–104] To evaluate temperature dependence, we repeated our MD simulation protocol at 310 K, 315 K, 325 K, and 330 K. We projected these trajectories into the same latent space using SRV encoders trained on WT data and re-trained independent MDN propagators to learn the dynamics at each temperature. We repeated this procedure over ion concentrations 25 mM, 50 mM, 200 mM, and 400 mM while holding temperature fixed at 320 K. Given that only the propagator was re-trained, we conserved 50% of training cost compared to training the full LSS pipeline. As in the previous section, 10 $\times$ 200 μs synthetic trajectories were generated and decoded into configuration space using cWGAN decoders trained on the WT

data.

We find that the thermodynamics of synthetic trajectories are in strong agreement with training data under our transfer learning procedure. In Figure 10A we show the fraction of hybridized duplex and folded hairpin as a function of temperature. We observe a distinct melting transition over the 20 K temperature range where the system shifts from mostly hybridized to dissociated at an approximate melting temperature of 317 K. There is a wider melting transition for hairpin DNA, which is expected given lower entropic favorability at higher temperatures.[105] The synthetic data quantitatively reproduces these trends within error bars and also reduces the statistical uncertainty, especially at low temperatures where half of training trajectories contain only a single (de)hybridization event. For example, at 310 K duplex population in $(78 \pm 18)\%$ for training data and $(87 \pm 6)\%$ for synthetic data and the hairpin population is $(53 \pm 10)\%$ for training data and $(49 \pm 3)\%$. Figure 10B shows training and synthetic thermodynamics as a function of ion concentration. At low sodium concentration, there is minimal screening between negatively charged DNA strands, and we observe low duplex population in both training $(7 \pm 7)\%$ and synthetic $(7 \pm 3)\%$ trajectories. Duplex populations increase asymptotically with ion concentration to $(71 \pm 17)\%$ for training data and $(80 \pm 4)\%$ for synthetic data at 400 mM. The population of folded hairpins increase slightly with ion concentration, however less than might be expected given previous single-molecule studies.[105] Synthetic trajectories reproduce these trends extremely well, although not with quite the same quantitative accuracy as the temperature comparison. This may indicate that ion concentrations induce changes to hairpin structures that are less well represented within the shared SRV latent space or previously experienced by the cWGAN decoder.

We next test whether our Multi-LSS transfer learning procedure reproduces meaningful kinetics across temperature and ion concentration. Kinetics can be expressed in terms of k_{on} and k_{off} rates which are frequently monitored in single-molecule studies[102,105] and nanotechology applications such as DNA-PAINT.[106,107] We calculated rates by collecting all

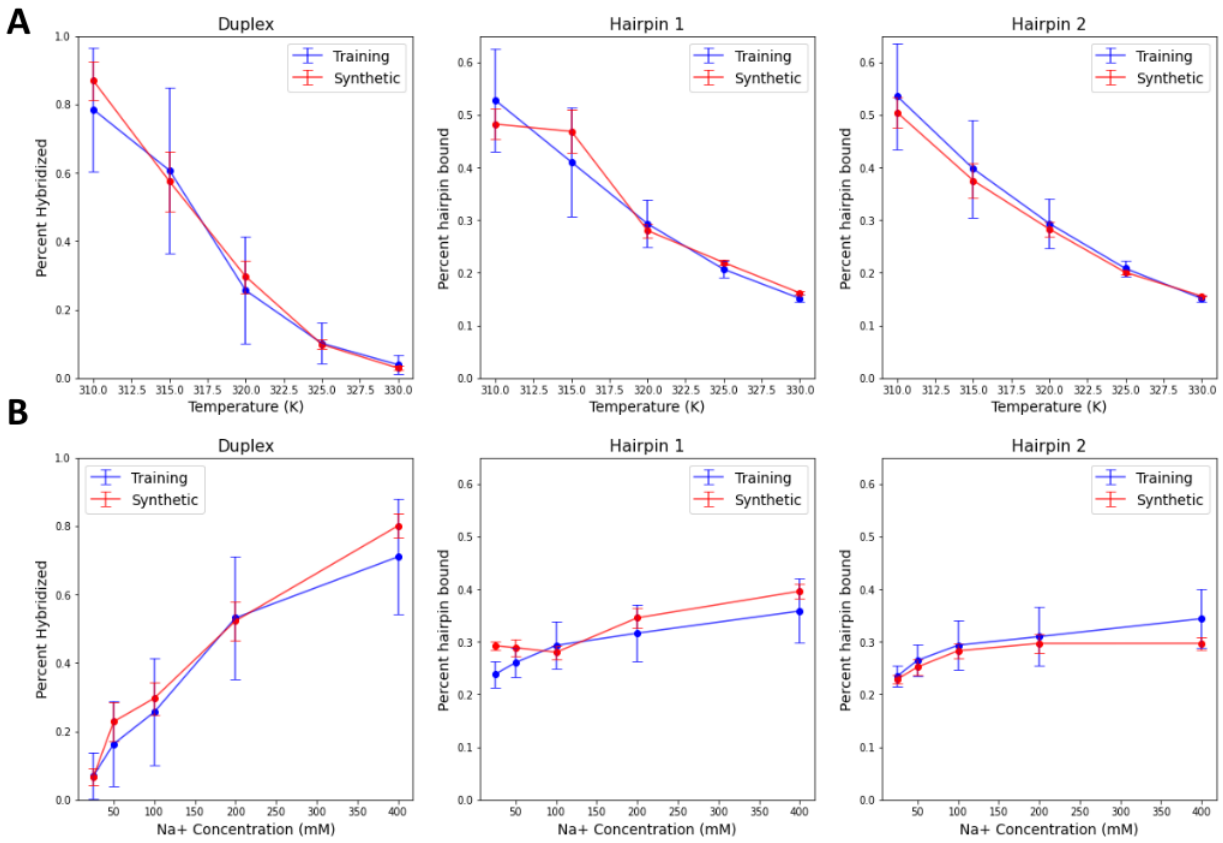

Figure 10: Comparison of thermodynamic predictions of temperature and salt concentration transferability of DNA Multi-LSS model. All synthetic data were generated using an SRV encoder and cWGAN decoder trained over the wild type conditions of 320 K and 100 mM NaCl; only the MDN propagator was retrained over simulation data collected under the new conditions. Hybridized duplex and folded hairpin fractions as a function of A) temperature at 100 mM NaCl and B) ion concentration at 320 K. Error bars represent the standard deviation across 10 independent trajectories.

hybridized/dissociated dwell times and fitting a single exponential distribution to determine k_{on} and k_{off} (Figure 11). Uncertainties were determined via a bootstrapping procedure where dwell times were randomly sampled with replacement and the model refitted. As expected for an activated process, we observe an exponential increase in k_{off} with an increase temperature, and a minimal impact on k_{on} [106,108] (Figure 11A). We see very close agreement for synthetic k_{off} rates as a function of temperature, but find that k_{on} rates are, although still very good, slightly faster on average than the training data. We see closer agreement

for k_{on} rates as a function of ion concentration, where association increases approximately linearly in agreement with single-molecule experiments over a similar concentration range[105] (Figure 11B). We calculated analogous rates – k_{fold} and k_{unfold} – for the hairpin folding process by using a cutoff along EE1 and EE2 to discretize hairpin trajectories into folded and unfolded regions. In agreement with the training data, we observe that synthetic k_{unfold} rates increase with temperature and decrease with ion concentration (Figure 11A,B). Synthetic k_{fold} rates qualitatively follow training rates, however, for both temperature and ion concentration we observe systematically higher synthetic k_{fold} values that lie slightly outside training uncertainties. We attribute this in part to a higher degree of structural variance in synthetic hairpin states that may lead to more frequent re-crossing of the binding cutoff.

4 Conclusions

In this work, we have applied our previously introduced molecular latent space simulators (LSS)[10] to a PROTAC ternary complex and a two-strand DNA system that can undergo both hybridization and hairpin folding. The former application demonstrates the viability of constructing LSS models of large biomolecular complexes of biomedical relevance from short, discontinuous training data generated by distributed computing. The latter application required the development of the Multi-LSS framework as a generalized LSS approach applicable to multi-molecular systems. In both cases, the trained LSS models can generate ultra-long trajectories of realistic molecular configurations with low computational overhead. Thermodynamics and kinetics of both systems were well reproduced, and statistical uncertainties are reduced by generating longer trajectories containing more conformational transitions than are available in the training data.

For the PROTAC system, we show that the learned latent encoding provides valuable insights into structural changes and dynamics associated with degradation efficiency. We hypothesize that latent coordinates can be leveraged as collective variables for the evalua-

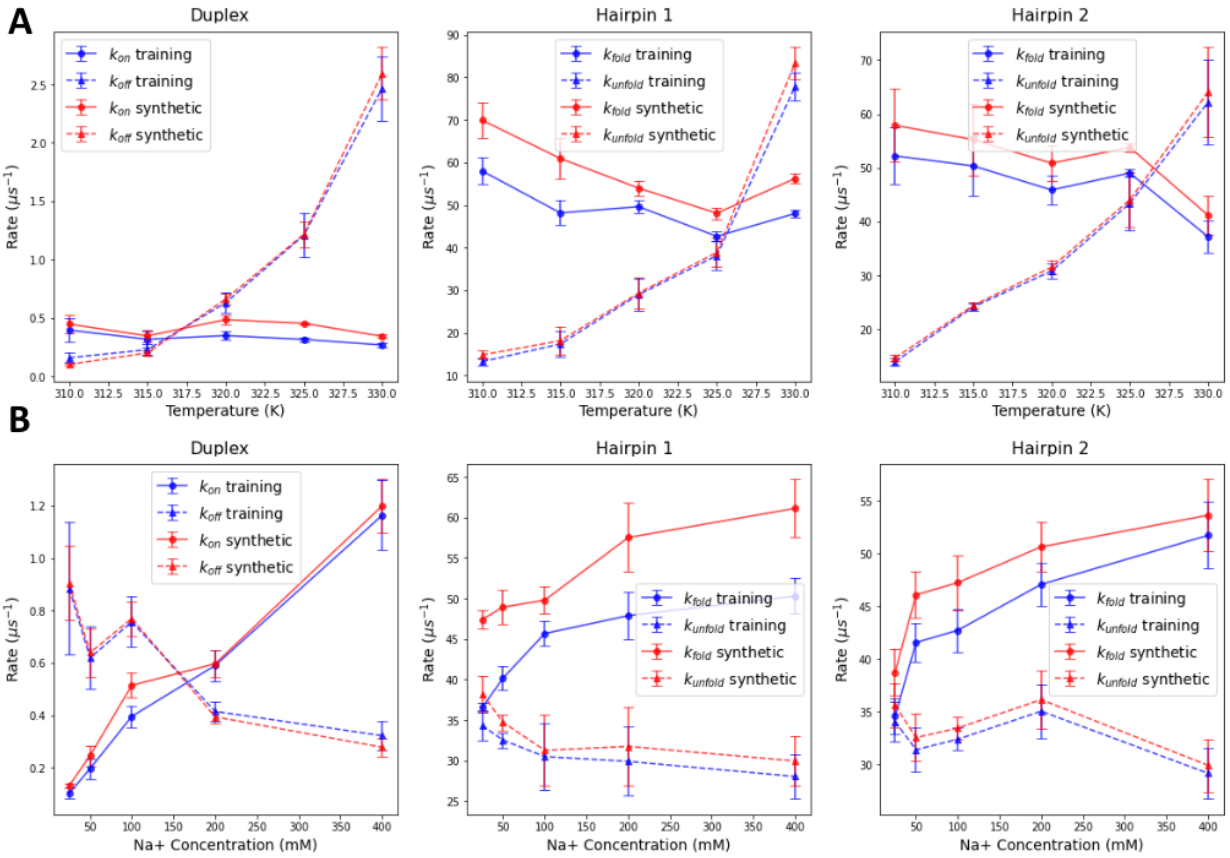

Figure 11: Comparison of kinetics predictions of temperature and salt concentration transferability of DNA Multi-LSS model. All synthetic data were generated using an SRV encoder and cWGAN decoder trained over the wild type conditions of 320 K and 100 mM NaCl; only the MDN propagator was retrained over simulation data collected under the new conditions. Hybridization and folding rates as a function of A) temperature at 100 mM NaCl and B) ion concentration at 320 K. Rates determined by fitting single exponential to empirical dwell time distributions. Error bars show standard deviation over 100 bootstrap samples.

tion of candidate PROTAC molecules, and furnish good coordinates for putative enhanced sampling calculations and high throughput virtual screening (HTVS) campaigns for rational *in silico* design of PROTACs with elevated degradation efficiency. We find that training on many short trajectories poses some limitations to the accurate estimation of timescales associated with transitions that are sparsely sampled or unsampled within the training data. However, this represents a challenge for any kinetic modeling technique including MSMs, and we are encouraged by the good kinetic and thermodynamic agreement for those states

and transitions that are sufficiently sampled in the training trajectories.

For the DNA system, the Multi-LSS approach introduces multiple encoders and decoders to capture the leading modes of the global system as well as of the individual sub-systems. This avoids the pitfalls associated with the exponential increase in the number of states and complexity of the transition matrix associated with the approximately independent dynamical evolution of the constituent subsystems in the non-interacting regime. Synthetic configurations can then be reproduced by decoding and re-orienting each individual strand or, for interacting configurations, by decoding of the global system configuration. We demonstrated this approach for a two-molecule system, but it is, in principle, scalable to systems containing arbitrary numbers of molecules. The model produces synthetic trajectories that substantially reduce both thermodynamic and kinetic statistical uncertainties associated with duplex hybridization and hairpin folding. This increase in precision becomes particularly important when folding and hybridization events are sparse, such as at low temperature or low ion concentration. Furthermore, we show that our encoder and decoder are transferable across a range of simulations conditions and that training efficiency can be enhanced by re-training the propagator alone. Although outside the scope of this work, we hypothesize that a propagator may be trained to interpolate or extrapolate over a tunable system parameter (e.g., temperature, pressure, salt concentration) in order produce meaningful synthetic trajectories under new thermodynamic states of interest without requiring the collection of new training data.

Although the LSS can push statistical uncertainties towards zero, we emphasize that systematic uncertainties present in the training data persist within the trained LSS model. The most common sources of systematic errors are approximations in the molecular force field and the presence of unsampled/undersampled states and transitions in the training trajectories. As an example of the first of these systematic errors, the artificial acceleration of hybridization dynamics associated with the 3SPN.2 coarse-grained model[60] are internalized by the LSS to produce rate estimates consistent with the training data but faster than what

might be observed experimentally or in all-atom calculations. Such systematic errors can, of course, only be ameliorated by the use of higher accuracy models in generation of the training data. As an example of the second of these systematic errors, poor sampling of particular transitions between metastable states of the PROTAC ternary complex along ψ_0 did not prevent the LSS model from learning a fully-connected long-time dynamical model, but it was compelled to "hallucinate" transition paths and timescales without a good ground-truth reference. Such systematic errors can be quite readily engaged by adaptive sampling strategies that determine where to collect new training data to better sample undersampled states and transitions. The LSS framework is particularly well-suited to adaptive sampling since the decoder can readily furnish molecular configurations to initialize new training simulations near the transition state of an undersampled transition or by extrapolating into unexplored regions of latent space. Cross validation protocols and an assessment of the change in the learned SRV modes, MDN transition density elements, and cWGAN reconstruction accuracy can then provide an internal assessment of when sufficient data has been collected and the training data are sufficiently rich and representative that the trained LSS model ceases to change with the addition of more training data.

In future work, we plan to continue to refine and extend the Multi-LSS paradigm to larger and more complex systems such as molecular self-assembly, to incorporate adaptive sampling strategies, and to explore the use of dynamical reweighting to enable the use of biased trajectories within model training.[109–111]

Conflict of Interest Statement

A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Applications 16/887,710 and 17/642,582, US Provisional Patent Applications 62/853,919, 62/900,420, 63/314,898, and 63/479,378 and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466.

Supporting Information

Robustness of MDN propagator to time-reversed training data; details on the selection of decoder cutoff distances; structural validation of synthetic DNA trajectories; additional MSM model details including Chapman-Kolmogorov (CK) tests, implied time scale plots, and mean-first passage times; four movies illustrating conformational changes to the PROTAC ternary complex in the leading SRV modes.

Author Information

Michael S. Jones Pritzker School of Molecular Engineering, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

Zachary A. McDargh Absci AI Research Lab, 152 West 57th St, New York, New York 10019, USA

Rafal P. Wiewiora Psivant Therapeutics, 451 D Street, Boston, Massachusetts 02210, USA

Jesus A. Izaguirre Atommap Corporation, 447 Broadway 2nd fl 187, New York, NY 10013, USA

Huafeng Xu Atommap Corporation, 447 Broadway 2nd fl 187, New York, NY 10013, USA

Andrew L. Ferguson Pritzker School of Molecular Engineering, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

email: andrewferguson@uchicago.edu

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. CHE-2152521. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on

the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629.

References

(1) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov state models but were afraid to ask. *Methods* **2010**, *52*, 99–105.

(2) Husic, B. E.; Pande, V. S. Markov state models: From an art to a science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

(3) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101.

(4) Harrigan, M. P.; Pande, V. S. Landmark Kernel tICA for conformational dynamics. *bioRxiv* **2017**, 123752.

(5) Chen, W.; Sidky, H.; Ferguson, A. L. Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets. *J. Chem. Phys.* **2019**, *150*, 214114.

(6) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 1–11.

(7) Sidky, H.; Chen, W.; Ferguson, A. L. Molecular latent space simulators. *Chem. Sci.* **2020**, *11*, 9459–9467.

(8) Suárez, E.; Wiewiora, R. P.; Wehmeyer, C.; Noé, F.; Chodera, J. D.; Zuckerman, D. M. What Markov state models can and cannot do: Correlation versus path-based observables in protein-folding models. *J. Chem. Theory Comput.* **2021**, *17*, 3119–3133.

(9) Hempel, T.; Olsson, S.; Noé, F. Markov field models: Scaling molecular kinetics approaches to large molecular machines. *Curr. Opin. Struct. Biol.* **2022**, *77*, 102458.

(10) Sidky, H.; Chen, W.; Ferguson, A. L. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.* **2020**, *118*, e1737742.

(11) Kevrekidis, I. G.; Gear, C. W.; Hyman, J. M.; Kevrekidis, P. G.; Runborg, O.; Theodoropoulos, C. Equation-free, coarse-grained multiscale computation: Enabling mocroscopic simulators to perform system-level analysis. *Commun. Math. Sci.* **2003**, *1*, 715–762.

(12) Kevrekidis, I. G.; Gear, C. W.; Hummer, G. Equation-free: The computer-aided analysis of complex multiscale systems. *AIChE J.* **2004**, *50*, 1346–1355.

(13) Kevrekidis, I. G.; Samaey, G. Equation-free multiscale computation: Algorithms and applications. *Annu. Rev. Phys. Chem.* **2009**, *60*, 321–344.

(14) Mori, H. Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.* **1965**, *33*, 423–455.

(15) Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **1973**, *9*, 215–220.

(16) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: Oxford, 2001.

(17) Risken, H.; Frank, T. *The Fokker-Planck Equation: Methods of Solution and Applications*, 2nd ed.; Springer Verlag: Berlin Heidelberg New York, 2012.

(18) Chen, W.; Sidky, H.; Ferguson, A. L. Capabilities and limitations of time-lagged autoencoders for slow mode discovery in dynamical systems. *J. Chem. Phys.* **2019**, *151*, 1–16.

(19) Wu, H.; Mardt, A.; Pasquali, L.; Noe, F.; Deep Generative Markov State Models, In *Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; 2018; Vol. 31.

(20) Fu, X.; Xie, T.; Rebello, N. J.; Olsen, B. D.; Jaakkola, T. Simulate time-integrated

coarse-grained molecular dynamics with geometric machine learning. *arXiv preprint arXiv:2204.10348* **2022**,

(21) Lin, K.; Peng, J.; Xu, C.; Gu, F. L.; Lan, Z. Realization of the trajectory propagation in the MM-SQC dynamics by using machine learning. *arXiv preprint arXiv:2207.05556* **2022**,

(22) Tsai, S.-T.; Kuo, E.-J.; Tiwary, P. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nat. Commun.* **2020**, *11*, 5115.

(23) Vlachas, P. R.; Zavadlav, J.; Praprotnik, M.; Koumoutsakos, P. Accelerated simulations of molecular systems through learning of effective dynamics. *J. Chem. Theory Comput.* **2022**, *18*, 538–549.

(24) Dixon, T.; MacPherson, D.; Mostofian, B.; Dauzhenka, T.; Lotz, S.; McGee, D.; Shechter, S.; Shrestha, U.; Wiewiora, R.; McDargh, Z. A.; et al., Predicting the structural basis of targeted protein degradation by integrating molecular dynamics simulations with structural mass spectrometry. *Nat. Commun.* **2022**, *13*, 5884.

(25) Sun, X.; Gao, H.; Yang, Y.; He, M.; Wu, Y.; Song, Y.; Tong, Y.; Rao, Y. PROTACs: Great opportunities for academia and industry. *Curr. Signal Transduct. Ther.* **2019**, *4*, 64.

(26) Zeng, S.; Huang, W.; Zheng, X.; Liyan cheng,; Zhang, Z.; Wang, J.; Shen, Z. Proteolysis targeting chimera (PROTAC) in drug discovery paradigm: Recent progress and future challenges. *Eur. J. Med. Chem.* **2021**, *210*, 112981.

(27) Sakamoto, K. M.; Kim, K. B.; Kumagai, A.; Mercurio, F.; Crews, C. M.; Deshaies, R. J. Protacs: Chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 8554–8559.

(28) Békés, M.; Langley, D. R.; Crews, C. M. PROTAC targeted protein degraders: The past is prologue. *Nat. Rev. Drug Discov* **2022**, *21*, 181–200.

(29) Drummond, M. L.; Williams, C. I. In silico modeling of PROTAC-mediated ternary complexes: Validation and application. *J. Chem. Inf. Model.* **2019**, *59*, 1634–1644.

(30) Bai, N.; Miller, S. A.; Andrianov, G. V.; Yates, M.; Kirubakaran, P.; Karanicolas, J. Rationalizing PROTAC-mediated ternary complex formation using Rosetta. *J. Chem. Inf. Model.* **2021**, *61*, 1368–1382.

(31) Bai, N.; Riching, K. M.; Makaju, A.; Wu, H.; Acker, T. M.; Ou, S.-C.; Zhang, Y.; Shen, X.; Bulloch, D. N.; Rui, H.; et al., Modeling the CRL4A ligase complex to predict target protein ubiquitination induced by cereblon-recruiting PROTACs. *J. Biol. Chem.* **2022**, *298*, 101653.

(32) Beberg, A. L.; Ensign, D. L.; Jayachandran, G.; Khaliq, S.; Pande, V. S. Folding@home: Lessons from eight years of volunteer distributed computing. *IPDPS 2009 - Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium* **2009**, 1–8.

(33) Mardt, A.; Hempel, T.; Clementi, C.; Noé, F. Deep learning to decompose macromolecules into independent Markovian domains. *Nat. Commun.* **2022**, *13*, 7101.

(34) Olsson, S.; Noé, F. Dynamic graphical models of molecular kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 15001–15006.

(35) Dibak, M.; Del Razo, M. J.; De Sancho, D.; Schütte, C.; Noé, F. MSM/RD: Coupling Markov state models of molecular kinetics with reaction-diffusion simulations. *J. Chem. Phys.* **2018**, *148*, 214107.

(36) Del Razo, M. J.; Dibak, M.; Schütte, C.; Noé, F. Multiscale molecular kinetics by

coupling Markov state models and reaction-diffusion dynamics. *J. Chem. Phys.* **2021**, *155*, 124109.

(37) Schreck, J. S.; Ouldridge, T. E.; Romano, F.; Šulc, P.; Shaw, L. P.; Louis, A. A.; Doye, J. P. DNA hairpins destabilize duplexes primarily by promoting melting rather than by inhibiting hybridization. *Nucleic Acids Res.* **2015**, *43*, 6181–6190.

(38) Hata, H.; Kitajima, T.; Suyama, A. Influence of thermodynamically unfavorable secondary structures on DNA hybridization kinetics. *Nucleic Acids Res.* **2018**, *46*, 782–791.

(39) Gao, Y.; Wolf, L. K.; Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: A solution versus surface comparison. *Nucleic Acids Res.* **2006**, *34*, 3370–3377.

(40) Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. Influence of secondary structure on kinetics and reaction mechanism of DNA hybridization. *Nucleic Acids Res.* **2007**, *35*, 2875–2884.

(41) Ding, F.; Cocco, S.; Raj, S.; Manosas, M.; Nguyen, T. T. T.; Spiering, M. M.; Bensimon, D.; Allemand, J.-F.; Croquette, V. Displacement and dissociation of oligonucleotides during DNA hairpin closure under strain. *Nucleic Acids Res.* **2022**, *50*, 12082–12093.

(42) Green, S. J.; Lubrich, D.; Turberfield, A. J. DNA hairpins: Fuel for autonomous DNA devices. *Biophys. J.* **2006**, *91*, 2966–2975.

(43) Seeman, N. C.; Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mater.* **2017**, *3*, 1–23.

(44) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(46) Van Gunsteren, W. F.; Berendsen, H. J. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1988**, *1*, 173–185.

(47) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(48) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(49) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(50) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(51) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

(52) Krzanowski, W. *Principles of Multivariate Analysis*; Oxford University Press, 2000.

(53) Grice, J. W.; Assad, K. K. Generalized Procrustes analysis: A tool for exploring aggregates and persons. *Appl. Multivar. Res.* **2009**, *13*, 93.

(54) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.

(55) Rentzeperis, D.; Shikiya, R.; Maiti, S.; Ho, J.; Marky, L. A. Folding of intramolecular DNA hairpin loops: Enthalpy-entropy compensations and hydration contributions. *J. Phys. Chem. B.* **2002**, *106*, 9945–9950.

(56) Reiling, C.; Khutsishvili, I.; Huang, K.; Marky, L. A. Loop contributions to the folding thermodynamics of DNA straight hairpin loops and pseudoknots. *J. Phys. Chem. A.* **2015**, *119*, 1939–1946.

(57) Mu, Z.-C.; Tan, Y.-L.; Zhang, B.-G.; Liu, J.; Shi, Y.-Z. Ab initio predictions for 3D structure and stability of single-and double-stranded DNAs in ion solutions. *PLoS Comput. Biol.* **2022**, *18*, e1010501.

(58) Hinckley, D. M.; Freeman, G. S.; Whitmer, J. K.; De Pablo, J. J. An Experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys.* **2013**, *139*, 1–17.

(59) Hinckley, D. M.; Lequieu, J. P.; de Pablo, J. J. Coarse-grained modeling of DNA oligomer hybridization: length, sequence, and salt effects. *J. Chem. Phys.* **2014**, *141*, 035102.

(60) Jones, M. S.; Ashwood, B.; Tokmakoff, A.; Ferguson, A. L. Determining sequence-dependent DNA oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, Markov state models, and infrared spectroscopy. *J. Am. Chem. Soc.* **2021**, *143*, 17395–17411.

(61) Lu, W.; Bueno, C.; Schafer, N. P.; Moller, J.; Jin, S.; Chen, X.; Chen, M.; Gu, X.; de Pablo, J. J.; Wolynes, P. G. OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *PLoS Comput. Biol.* **2021**, *17*, e1008308.

(62) Tan, C.; Takada, S. Dynamic and structural modeling of the specificity in protein-DNA

interactions guided by binding assay and structure data. *J. Chem. Theory Comput.* **2018**, *14*, 3877–3889.

(63) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1997**, *117*, 1–42.

(64) Dunweg, B.; Paul, W. Brownian dynamics simulations without Gaussian random numbers. *Int. J. Mod. Phys. C.* **1991**, *2*, 817–27.

(65) Bussi, G.; Parrinello, M. Accurate sampling using Langevin dynamics. *Phys. Rev. E.* **2007**, *75*, 056707.

(66) Nkodo, A. E.; Garnier, J. M.; Tinland, B.; Ren, H.; Desruisseaux, C.; McCormick, L. C.; Drouin, G.; Slater, G. W. Diffusion coefficient of DNA molecules during free solution Electrophor. *Electrophor.* **2001**, *22*, 2424–2432.

(67) Debye, P.; E., H. Zur Theorie der Elektrolyte. *Phys. Z.,* **1923**, 185–206.

(68) Schneider, T.; Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B.* **1978**, *17*, 1302–1322.

(69) Arnott, S.; Smith, P. J. C.; Chandrasekaran, *CRC Handbook of Biochemistry and Molecular Biology*; CRC Press, 1976; pp 411–422.

(70) Schönemann, P. H. A generalized solution of the orthogonal Procrustes problem. *Psychometrika* **1966**, *31*, 1–10.

(71) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272.

(72) Noé, F.; Nüske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.* **2013**, *11*, 635–655.

(73) Klus, S.; Nüske, F.; Koltai, P.; Wu, H.; Kevrekidis, I.; Schütte, C.; Noé, F. Data-driven model reduction and transfer operator approximation. *J. Nonlinear Sci.* **2018**, *28*, 985–1010.

(74) Wu, H.; Noé, F. Variational ppproach for learning Markov processes from time series data. *J. Nonlinear Sci.* **2020**, *30*, 23–66.

(75) Nuske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S.; Noé, F. Variational approach to molecular kinetics. *J. Chem. Theory Comput.* **2014**, *10*, 1739–1752.

(76) Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K.; Deep canonical correlation analysis, In *Proceedings of the 30th International Conference on Machine Learning*; Dasgupta, S., McAllester, D., Eds.; 2013; Vol. 28; pp 1247–1255.

(77) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(78) Bonati, L.; Piccini, G.; Parrinello, M. Deep learning the slow modes for rare events sampling. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2113533118.

(79) Kingma, D. P.; Ba, J. L. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* **2015**, 1–15.

(80) Bishop, C. M. Mixture density networks. *Neural Computing Research Group Report* **1994**, 171–198.

(81) Bishop, C. M. Pattern Recognition and Machine Learning. *Springer-Verlag New York, Inc., Secaucus, NJ, USA* **2006**, *58*, 9.

(82) Noé, F. Machine learning for molecular dynamics on long timescales. *Lecture Notes in Physics* **2020**, *968*, 331–372.

(83) Pathak, J.; Hunt, B.; Girvan, M.; Lu, Z.; Ott, E. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Phys. Rev. Lett.* **2018**, *120*, 24102.

(84) Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. *34th International Conference on Machine Learning, ICML 2017* **2017**, *1*, 298–321.

(85) Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. C.; Improved Training of Wasserstein GANs, In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; 2017; Vol. 30.

(86) Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* **2014**,

(87) Farnaby, W.; Koegl, M.; Roy, M. J.; Whitworth, C.; Diers, E.; Trainor, N.; Zollman, D.; Steurer, S.; Karolyi-Oezguer, J.; Riedmueller, C.; et al., BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. *Nat. Chem. Biol.* **2019**, *15*, 672–680.

(88) Nguyen, H. C.; Wang, W.; Xiong, Y. Cullin-RING E3 ubiquitin ligases: Bridges to destruction. *Macromolecular Protein Complexes: Structure and Function* **2017**, 323–347.

(89) Huang, H.-T.; Dobrovolsky, D.; Paulk, J.; Yang, G.; Weisberg, E. L.; Doctor, Z. M.; Buckley, D. L.; Cho, J.-H.; Ko, E.; Jang, J.; et al., A chemoproteomic approach to query the degradable kinome using a multi-kinase degrader. *Cell Chem. Biol.* **2018**, *25*, 88–99.

(90) Schiemer, J.; Horst, R.; Meng, Y.; Montgomery, J. I.; Xu, Y.; Feng, X.; Borzilleri, K.; Uccello, D. P.; Leverett, C.; Brown, S.; et al., Snapshots and ensembles of BTK and cIAP1 protein degrader ternary complexes. *Nat. Chem. Biol.* **2021**, *17*.

(91) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079–2102.

(92) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *J. Chem. Phys.* **2018**, *149*, 072312.

(93) Sultan, M. M.; Pande, V. S. TICA-metadynamics: Accelerating metadynamics by using kinetically selected collective variables. *J. Chem. Theory Comput.* **2017**, *13*, 2440–2447.

(94) Stasi, M.; Monferrer, A.; Babl, L.; Wunnava, S.; Dirscherl, C.; Braun, D.; Schwille, P.; Dietz, H.; Boekhoven, J. Regulating the dynamic folding of a DNA hairpin at the expense of a small, molecular fuel. *ChemRxiv preprint: 10.26434/chemrxiv-2021-w299m* **2021**, 1–7.

(95) Irmisch, P.; Ouldridge, T. E.; Seidel, R. Modeling DNA-strand displacement reactions in the presence of base-pair mismatches. *J. Am. Chem. Soc.* **2020**, *142*, 11451–11463.

(96) Lapteva, A. P.; Sarraf, N.; Qian, L. DNA strand-displacement temporal logic circuits. *J. Am. Chem. Soc.* **2022**, *144*, 12443–12449.

(97) Kennedy, T.; Pearce, C.; Thachuk, C.; Fast and robust strand displacement cascades via systematic design strategies, In *28th International Conference on DNA Computing and Molecular Programming (DNA 28)*; Ouldridge, T. E., Wickham, S. F. J., Eds.; 2022; Vol. 238.

(98) Sidky, H.; Chen, W.; Ferguson, A. L. High-resolution Markov state models for the dynamics of Trp-cage miniprotein constructed over slow folding modes identified by state-free reversible VAMPnets. *J. Phys. Chem. B.* **2019**, *123*, 7999–8009.

(99) Scherer, M. K.; Trendelkamp-schroer, B.; Paul, F.; Pe, G.; Ho, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-h.; Noe, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

(100) Röblitz, S.; Weber, M. Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv. Data Anal. Classif.* **2013**, *7*, 147–179.

(101) Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable neural networks for enhanced sampling of protein dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 1887–1894.

(102) Nicholson, D. A.; Jia, B.; Nesbitt, D. J. Measuring excess heat capacities of deoxyribonucleic acid (DNA) folding at the single-molecule level. *J. Phys. Chem. B.* **2021**, *125*, 9719–9726.

(103) Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A. Effects of sodium ions on DNA duplex oligomers: Improved predictions of melting temperatures. *Biochem.* **2004**, *43*, 3537–3554.

(104) Lipfert, J.; Doniach, S.; Das, R.; Herschlag, D. Understanding nucleic acid—ion interactions. *Annu. Rev. Biochem.* **2014**, *83*, 813–841.

(105) Tsukanov, R.; Tomov, T. E.; Masoud, R.; Drory, H.; Plavner, N.; Liber, M.; Nir, E. Detailed study of DNA hairpin dynamics using single-molecule fluorescence assisted by DNA origami. *J. Phys. Chem. B.* **2013**, *117*, 11932–11942.

(106) Jungmann, R.; Steinhauer, C.; Scheible, M.; Kuzyk, A.; Tinnefeld, P.; Simmel, F. C. Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. *Nano Lett.* **2010**, *10*, 4756–4761.

(107) Schnitzbauer, J.; Strauss, M. T.; Schlichthaerle, T.; Schueder, F.; Jungmann, R. Super-resolution Microscopy with DNA-PAINT. *Nat. Protoc.* **2017**, *12*, 1198–1228.

(108) Andrews, R. DNA hybridisation kinetics using single-molecule fluorescence imaging. *Essays Biochem.* **2021**, *65*, 27–36.

(109) Donati, L.; Hartmann, C.; Keller, B. G. Girsanov reweighting for path ensembles and Markov state models. *J. Chem. Phys.* **2017**, *146*, 244112.

(110) Donati, L.; Keller, B. G. Girsanov reweighting for metadynamics simulations. *J. Chem. Phys.* **2018**, *149*, 072335.

(111) Shmilovich, K.; Ferguson, A. L. Girsanov Reweighting Enhanced Sampling Technique (GREST): On-the-Fly Data-Driven Discovery of and Enhanced Sampling in Slow Collective Variables. *J. Phys. Chem. A* **2023**, *127*, 3497–3517.

Graphical TOC Entry